IMF Working Paper

FinTech in Financial Inclusion
Machine Learning Applications in Assessing Credit Risk

By Majid Bazarbash

INTERNATIONAL MONETARY FUND

WP/19/109

IMF Working Paper

Monetary and Capital Markets Department

FinTech in Financial Inclusion:
Machine Learning Applications in Assessing Credit Risk

Prepared by Majid Bazarbash[*]

Authorized for distribution by Ulric Eriksson von Allmen

May 2019

Abstract

Recent advances in digital technology and big data have allowed FinTech (financial technology) lending to emerge as a potentially promising solution to reduce the cost of credit and increase financial inclusion. However, machine learning (ML) methods that lie at the heart of FinTech credit have remained largely a black box for the nontechnical audience. This paper contributes to the literature by discussing potential strengths and weaknesses of ML-based credit assessment through (1) presenting core ideas and the most common techniques in ML for the nontechnical audience; and (2) discussing the fundamental challenges in credit risk analysis. FinTech credit has the potential to enhance financial inclusion and outperform traditional credit scoring by (1) leveraging nontraditional data sources to improve the assessment of the borrower's track record; (2) appraising collateral value; (3) forecasting income prospects; and (4) predicting changes in general conditions. However, because of the central role of data in ML-based analysis, data relevance should be ensured, especially in situations when a deep structural change occurs, when borrowers could counterfeit certain indicators, and when agency problems arising from information asymmetry could not be resolved. To avoid digital financial exclusion and redlining, variables that trigger discrimination should not be used to assess credit rating.

Contents

ACRONYMS

4GL	4th Generation Programming Language	**LTV**	Loan-to-Value Ratio
BIS	Bank of International Settlements	**ML**	Machine Learning
DBMS	Database Management Systems	**MSE**	Mean Squared Error
DTI	Debt-to-Income Ratio	**NN**	Neural Network
FSB	Financial Stability Board	**PD**	Probability of Default
GBDT	Gradient Boosting Decision Tree	**PDP**	Partial Dependent Plot
IMF	International Monetary Fund	**SQL**	Structured Query Language
LASSO	Least Absolute Shrinkage Selection Operator	**SVM**	Support Vector Machine
LGD	Loss Given Default	**VAR**	Vector Autoregressive Model

GLOSSARY

Five Cs of Credit	A system to assess credit risk of business borrowers based on five features of the borrower: character, capacity, capital structure, coverage, and conditions
Cross-Validation	A model validation technique for assessing how the results of a statistical model will generalize to an independent data set and can make predictions
Estimation Error	Average difference between the estimated value of the target variable and realized values. Common measures are the mean squared errors or mean absolute errors.
FinTech	Technologically enabled financial innovation that could result in new business models, applications, processes or products with an associated material effect on financial markets, financial institutions and provision of financial services
Generalization	The ability of a model to offer comparable in-sample and out-of-sample predictive performance
Machine Learning	A method of designing a sequence of actions to solve a problem that optimize automatically through experience and with limited or no human intervention.
Nonparametric Model	A data analysis models whose structure is not specified a priori, but is instead determined from data
Overfitting	The problem of a statistical model that offers high in-sample, fit but poor out-of-sample predictive performance

I. INTRODUCTION

One of the major challenges of low-income and emerging economies is the high cost of credit. The high cost of borrowing and credit rationing gives rise to financial exclusion of especially small borrowers such as small- and medium-size enterprises and households that play a significant role in the macroeconomy ([Sahay et al. (2015a)](#)). As a promising solution, modern technological advances have enabled new business models to employ modern data analysis techniques on big data and automate tasks to make credit decisions more efficiently. FinTech credit[1] promises to offer loans at a higher speed and lower cost, and therefore grant loans to a larger fraction of the population, resulting in elevated financial inclusion[2].

Policymakers and economists have recognized FinTech's potential to transform the financial system but have also raised concerns about the opaque nature of the way modern technology could create value and generate financial stability risks ([The Bali FinTech Agenda (2018)](#)). In particular, for assessing credit risk, FinTech companies rely heavily on machine learning (ML) techniques that have been developed in the computer science field. Economists who want to improve their understanding of financial inclusion in the new FinTech landscape need to understand the technology, its impact on credit extension, and the benefits and risks it carries.

This paper's contribution to the literature in financial inclusion and FinTech is that it provides a nontechnical analysis of how ML could enhance assessing credit risk of borrowers while describing the main strengths and weaknesses of ML-based lending. To this end, the paper takes two important steps. First, to address the finance aspect of FinTech, it discusses the fundamental challenges that arise when assessing credit risk, confronted by both FinTech and traditional lenders. Second, the paper takes ML out of the black box by going over key concepts and ideas in ML and reviewing the most common techniques in ML in a nontechnical and illustrative way. Building on these two steps, the paper then discusses what ML-based lending can and cannot do to enhance credit assessment and ways in which FinTech credit can contribute to responsible financial inclusion. Therefore, it takes a step toward drawing policy implications for using FinTech credit to promote financial inclusion.

To describe the underlying challenges in prudent lending, the paper draws on the credit risk modeling literature. Credit risk assessment is seen as estimating a loan's probability of default and loss given default while taking agency problems arising from information asymmetry into consideration. Practitioners often seek information for five broad attributes (capacity, capital structure, coverage, character, and conditions), the well-known five Cs of credit ratings. ML has the potential to improve the assessments of some of the five Cs of the credit rating process, and thereby improve credit screening more broadly. For instance, [Jagtiani & Lemieux (2017)](#) document superior performance of loans originated by Lending

---

[1] The use of machine learning for assessing credit risk is not confined to FinTech credit companies, but has been incorporated by some large banks in their credit rating process at least in form of the challenger model. Nonetheless, FinTech credit companies, including peer-to-peer lending platforms and balance sheet lenders, use ML as their major method for assessing credit risk.

[2] For a discussion of macroeconomic gains of improving financial inclusion, see [Sahay et al. (2015b)](#).

Club in the US compared to traditional lenders and find a more accurate risk pricing through ML-based credit analysis.

The paper provides a survey of the recent ML literature and discusses the several methods that have gained the highest popularity in practice. These models include decision trees, random forest, gradient boosting trees, support vector machines, and neural networks. The paper presents the fundamental ideas underpinning these methods, and discusses their strengths, weaknesses, and extensions. While the discussion avoids mathematical detail and computational methods involved in applying the algorithms, it provides sufficient background for a nontechnical learner to understand the models.

The core philosophy of ML is to apply potentially complicated algorithms that could be run by machines to learn patterns in data with the primary goal of making predictions. ML models are designed to analyze a large amount of information contained in data from various sources. ML models could identify patterns in the data that standard econometric models cannot[3]. Nonetheless, a major shortcoming of using complex algorithms is that, in general, these patterns cannot be readily communicated with analysts and verified against expert knowledge. Therefore, while ML could make remarkable use of available data for making more accurate predictions than traditional methods, it could generate misleading results when data relevance is questioned. For example, in situations where there are rapid structural changes in the environment that are not fully reflected in the data, a naïve use of ML models may overweight information prior to the structural change and, for instance, result in the underestimation of the credit risk of borrowers.

An evolving body of literature has emerged in recent years that studies various aspects of FinTech, trends in FinTech development, its economic impact, potential risks, and possible policy and regulatory implications (see, for example, contributions by Phillipon (2016), Berg et al. (2018), Buchak et al. (2018), Das (2018), Fuster et al. (2018), Gomber et al. (2018)[4], by IMF staff: He et al. (2017), Lukonga (2018), Berkmen et al. (2019), McNeill et al. (forthcoming) and The Bali FinTech Agenda (2018), FSB: (2017a, 2017b), CGSF & FSB (2017), BIS: Claessens et al. (2018), BIS (2018), and CGAP: Kaffenberger et al. (2018))[5]. These studies focus on the implications of FinTech, taking a rather broad view of ML. To facilitate a more in-depth understanding of the strengths and shortcomings of using ML for assessing credit risk and its implications for fostering financial inclusion, this paper provides a nontechnical primer to fundamental concepts in ML and some of the most common ML techniques.

A body of literature in economics presents studies of the applications of machine learning in various areas. For an extensive discussion of applications of machine learning models in economics, see NBER 2017 conference papers under "The Economics of Artificial

---

[3] See for instance, Lotterman et al. (2012), among others, who show superior performance of machine learning models, including support vector machines and neural networks relative to the typical linear models in predicting loss given default. For a recent example, see Bazarbash (2019).

[4] For a review of recent academic literature studying the use of digital technology in finance, see Gomber et al. (2017).

[5] The list is highly selective and not intended to be comprehensive.

Intelligence: An Agenda," particularly Athey (2018). Other similar studies are Varian (2014), Athey & Imbens (2015), and Mullainathan & Spiess (2017). This paper takes a less technical approach and discusses ML in the context of credit risk assessment, which is not the main focus of prior studies.

There has been a surge in recent years in the use of ML tools for estimating credit risk, especially since the establishment of Basel II, which called for development of internal credit rating models by banks, and since the global financial crisis. However, internal rating models based on the standard linear econometric approach have been generally shown to exhibit poor performance in forecasting loss given default (Altman and Hotchkiss (2010)). Studies of credit risk show that while ML models outperform traditional models, their performance depends on the specific ML model, the environment, and the sample used in the analysis (see Bastos (2010), Qi and Zhao (2011), Lotterman et al. (2012), Tows (2016), Bazarbash (2017), and Nazemi et al. (2018)). Motivated by these studies, this paper covers the most common and powerful ML methods applied for credit analysis.

The rest of the paper is organized as follows. Section II presents the main elements of prudent lending in the context of the five Cs of credit and agency problems. Section III discusses main ML tools and techniques. Building on these sections, Section IV analyzes strengths and weaknesses of ML-based lending in contrast to traditional methods. Section V includes concluding remarks.

II. THE FIVE Cs OF CREDIT RATING, PD, AND LGD

Lending decisions involve evaluating a borrower's credit risk, which consists of assessing the capacity and motivation of the borrower to repay the loan and the lender's protection against losses if the borrower defaults. The credit rating is summarized by two important metrics: the probability of default (PD) and the expected loss given default (LGD). The PD rating shows the likelihood of failing to repay the loan, and therefore depends on the income-generating capacity of the borrower during the loan lifecycle. The LGD measures the fraction of lost value relative to the outstanding balance at default, which is linked to the lender's protection against loss upon default.

Information to assess the PD and LGD of a loan comes from a wide range of sources. This section discusses the common framework the five Cs of credit, used by practitioners to group borrower information under five main features that are important for credit rating[6].The five Cs of credit rating is primarily used for business borrowers, but shares common attributes with credit rating of households. The five Cs of credit include capacity, capital structure, coverage, character, and conditions. An important way in which ML could enhance assessing credit risk is to improve evaluation of each of these features.

Capacity
The borrower's financial capacity shows the ability of the business to repay the loan, considering other senior obligations of the company and the company's profitability. A key indicator of capacity is the debt-to-income (DTI) ratio. A lower DTI in good times is often

---

[6] See Tirole (2010), Investopedia, and websites of Bank of America and Wells Fargo.

considered indicative of sufficient cash-generating capacity in downturns. However, the explanatory power of indicators like DTI for financial capacity depends crucially on the operational attributes of the underlying business.

Alternatively, a more fundamental analysis could directly model the future path of business income in terms of business performance incorporating factors such as:
- the business model;
- the infrastructure;
- key expertise;
- competitiveness;
- customer base;
- potential growth; and
- technological advantage.

Capital Structure
Capital structure reflects the composition of the company's liabilities. A higher capital ratio means a higher stake in the company by the business owners, which tends to provide a higher incentive for the managers to run the business profitably and delays default triggers due to a higher cushion provided by the capital.

Coverage
The effective value of pledged collateral and loan coverage indicates protection against loan losses in case of default. Loan coverage comes in two main forms: pledged collateral assets and loan guarantors. Like the capital ratio, higher security due to higher coverage provides the lender with more confidence in loan recovery in the event of default and reduces the LGD. An important challenge in evaluating coverage is the lack of a formal market for some assets, a situation that arises in particular for small business borrowers. Accordingly, the lender does not have a reference market price to use as a basis for estimating the price of such assets and project the future value evolution. In this case, a pricing model needs to be developed based on the specifics of each asset and possible implications of changes in the borrower's business on the asset's liquidation value.

Character
The borrower's character constitutes the track record for missed payments and default on previous debts, fraud, legal expenses, and basically any information implying questionable character of the business owner. Such information usually appears on the borrower's credit report and is used to generate a credit score (for example, the FICO score in the US) that banks predominantly use in credit assessment. A higher rating for character is often perceived as a manager with a higher sense of responsibility in running the business. The significance of character is particularly high in loans made to small companies, where the manager's role is pivotal in the success of the business.

Conditions
General conditions are beyond the company's control, but have cyclical influence on the loan performance. Conditions could include macroeconomic conditions (the business cycle) such as the growth rate, unemployment rate, inflation rate, interest rate, tariffs, regulations, the

financial cycle, industry-wide factors, and geography. Unlike borrower-specific features, such information remains the same for loans in the same industry and market. One way to think about the five Cs of credit rating is that "conditions" reflect systematic drivers while other Cs represent borrower-specific credit risk drivers. From a finance perspective, underlying conditions beyond the control of the business manager represent the nondiversifiable component of risk.

III. MACHINE LEARNING (ML) METHODS

This section presents some of the most prominent applications of machine learning that have proved successful in various applications and are used in credit risk analysis. It starts by reviewing fundamental concepts in machine learning and introducing the key concept of cross-validation in ML[7]. Next, it presents and discusses three classes of ML models, including tree-based models (decision trees, random forest, and gradient boosting trees), support vector machines, and neural networks[8].

A. Fundamental Concepts

What is Machine Learning?
At its core, ML takes advantage of the high computational power of "machines" to run (mostly complicated) algorithms to "learn" from data. Jordan and Mitchell (2015) define a learning problem as "improving some measure of performance when executing some task, through some type of training experience." The measure of performance is often defined as minimizing the overall error of estimation. The estimation error captures the average difference between the estimated value of the target variable and realized values. Common measures are the mean squared errors or mean absolute errors.

Machine learning models as nonparametric models
Machine learning models are largely nonparametric as opposed to econometric models that focus predominantly on parametric modeling[9]. Parametric models posit a functional form that links the outcome variable to explanatory variables based on a set of assumptions (or theory) that reflect the modeler's understanding of the relationship. For example, it may be assumed that price elasticities of demand are invariant with respect to the volume of demand and accordingly posit a log-linear function that could be estimated by an ordinary least square approach. Another example is to assume variables follow a mean-reverting and Markovian structure with an autocorrelation parameter invariant to time, and therefore assume an autoregressive model as a standard vector autoregressive model.

---

[7] The material in this section predominantly draws on standard ML textbooks by James et al. (2013) and Hastie et al. (2016) as well as a scikit-learn user guide.

[8] According to *Risk* magazine (July 3, 2018), the two most common ML techniques deployed in credit risk modeling are random forest and gradient boosting decision trees. Moreover, lenders are also experimenting with neural networks.

[9] A more detailed comparison of ML and econometric modeling approaches is given in the next section. ML also includes parametric models such as Least Absolute Selection and Shrinkage Operator (LASSO) and ridge regressions, used for dimension reduction when many features are included in the model.

Estimation of model parameters allows for quantifying and possibly testing the modeler's understanding of the functional form that is assumed to hold between the features and output variable. Other common examples of econometric models that are parametric are the ordinary least squares, logistic model, and structural models in economics. The advantages of parametric models are parsimony, interpretability, fast estimation speed, and less need for a large training dataset. However, the main weaknesses of parametric models are specification risk, limited complexity, and generally poor predictive power.

On the other hand, nonparametric models make minimal functional assumptions to gain more flexibility in learning from the data. All machine learning models discussed in this section fall under the class of nonparametric models. The main strength of nonparametric models is their high flexibility and therefore their high predictive performance, which increases when a larger sample is used to train the model [10]. The main concern in using nonparametric models is the risk of overfitting.

Cross-validation

The problem of overfitting is usually addressed in machine learning by a model assessment technique called cross-validation. In cross-validation, data are randomly partitioned into two samples called train and test samples. The information of features and responses in the train sample are used to estimate the model parameters, that is, "train" the model. The prediction error of the estimated model is measured by feeding the features of the test sample into the model and comparing the model's predictions with the actual values of the outcome variable. Because the train and test samples are independent of each other, the prediction error measures out-of-sample predictive power of the model, that is, prediction of response on a new observation [11].

In this context, a model with the problem of overfitting yields a low train error but a high test error. As a result, the model does not generalize well to make accurate predictions on the test sample and is over-parametrized.

k-fold cross-validation

In practice, a version of cross-validation called *k-fold cross-validation* is often used. Figure 1 illustrates a fivefold cross-validation for a sample with 120 observations. The objective is to randomly split the sample into five partitions with equal numbers of observations and use four partitions to train the model and the left-out partition to evaluate the prediction error based on, for instance, the mean squared error (MSE). Partitions 2 to 5 are used for training the model, and the prediction error is calculated on partition 1, call MSE_1. Similarly, leaving out partition 2, training the model on partitions 1, 3, 4, and 5, and evaluating test error yields MSE_2. The fivefold cross-validation prediction error is the average prediction errors on the five partitions, that is, $\frac{1}{5}\sum_{i=1}^{5} MSE_i$.

---

[10] As Russell and Norvig (2016) put it, "Nonparametric models are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing the right features."

[11] For example, in case of regressions, it is possible to obtain a negative R^2 for the goodness of fit on the test sample, because of including noisy variables among the features. The effect of noisy variables would not be captured by the R^2 of the model on the train sample, because the model minimizes the mean squared error using the actual values of the outcome variable.

Machine learning models often have hyper-parameters that govern the shape of the algorithm. Examples include the size of trees in the tree-based models, the learning rate in a support vector machine, and the number of layers and nodes in a neural network (discussed in more detail later in this section). These hyper-parameters are often calibrated to minimize the cross-validation prediction error. However, this approach implies that the information in the validation set is used in model estimation. As an alternative, the sample could be independently split into three subsamples: the train, test, and prediction. The train sample is used for estimating model parameters, the test sample is used to calibrate the hyper-parameters, and the prediction sample is used to evaluate the prediction error.

Figure 1. Fivefold Cross-Validation of a Sample with 120 Observations.

Randomly choose 24 obs	Randomly choose 24 obs Excluding partition 1	Randomly choose 24 obs Excluding partitions 1 & 2	Randomly choose 24 obs Excluding partitions 1,2 & 3	Exclude partitions 1,2,3 & 4
Partition 1	Partition 2	Partition 3	Partition 4	Partition 5
{3,6,…,111}	{5,10,…,120}	{1,2,…,110}	{4,8,…,98}	{7,9,…,115}
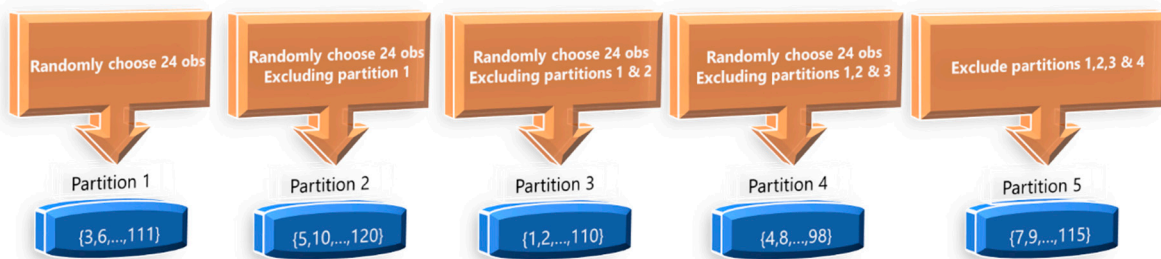
Note: The numbers in brackets show the indexes of the observations in the data. The average value of MSE of predictions on each partition, where that partition is left out in model estimation, represents the fivefold cross-validation MSE.

Supervised and unsupervised learning models
Another distinction often made between machine learning models is based on the availability of information for the outcome variable. When data for the outcome measure are available and the model is built based on labeled data, it is called a *supervised learning* model. The objective of a supervised learning model is to make predictions based on modeling the relationship between features and the outcome variable. For example, suppose a lender is interested in predicting the probability of default of auto loan applicants based on borrower attributes (for example, income, age, career, tenure, debt, etcetera), car features (for example, market price, gas usage, seating capacity, etcetera), and economic indicators (for example, interest rate, employment rate, inflation rate, etcetera). If in addition to these features, the dataset includes information about loan repayment behavior of past borrowers and their behavior can be considered representative of new applicants, the ML model is a supervised learning model. In this case, observed payment performance is used to train the ML model. Most of the models developed for credit analysis are supervised learning models.

On the other hand, if the dataset does not include values for the outcome variable, the ML model is an *unsupervised learning* model. Because in an unsupervised learning model information for the outcome variable is not available, the goal is to find similarities (such as modes or associations) among features and group observations based on the identified characteristics. In the example of credit rating auto loan applicants, suppose the dataset did not include repayment outcomes (called "unlabeled" data). In this case, believing that the features were relevant for explaining the loan performance behavior, unsupervised learning

models could be applied to cluster borrowers of similar features. If repayment information were partially available, it is assumed that the applicants in a cluster would behave similarly.[12]

To predict the credit risk using supervised learning models, in addition to having a labeled dataset, the relevancy condition should be satisfied. The relevancy condition requires the labeled dataset to be representative of the types of loans whose credit risk the model is intending to assess. For example, if new loans are believed to belong to a new line of business with vastly different risk properties, a supervised learning model could yield misleading predictions and is therefore inappropriate to use, even if the model displays high performance measures on the original dataset. Therefore, expert judgment is needed to ensure the dataset's relevance to the analysis, for example, by ruling out deep structural changes in the environment.

Classification versus regression models
Supervised machine learning models are classified based on the type of the outcome variable. The outcome variable is a categorical variable when possible outcomes of the variable are groups or labels, and the values of the variable indicates belonging to a group rather than reflecting a measurement, for instance, taking on 1 for female and 0 for male. The modeling exercise in which the outcome variable is categorical is called *classification*. Because default outcome is binary ("default" or "no default"), PD models are classification models. On the other hand, if a quantitative outcome variable is modeled, for example, LGD, the modeling exercise is called *regression*.

B.   Prominent Machine Learning Models

Some machine learning models in the literature have demonstrated good performance in a variety of predictive modeling exercises. The three most prominent classes of ML models presented in this section can be considered classic ML models with extensions that have attracted special interest in credit risk modeling. These models include: (1) tree-based models; (2) support vector machines; and (3) neural networks (NN). For easier exposition, I first explain the basic concept of each model and then provide an evaluation of the main strengths and weaknesses.

Tree-Based Models: Decision Trees, Random Forest, Gradient-Boosting Trees
Tree-based models in machine learning were inspired by if-then-else decision trees, which have been around for a long time in decision theory sciences. The if-then-else decision trees break down a large decision into a sequence of small decisions covering all possible outcomes in a hierarchical structure, where later decisions depend on former ones and the ultimate outcome is based on the sequence of all decisions in a branch. As a result, a clear connection is made between a set of possible decisions and their consequences, which have

---

[12] A third group of models is sometimes defined as a combination of supervised and unsupervised learning called *reinforcement learning*. Reinforcement learning starts with an unsupervised learning model based on unlabeled data, chooses an action with an impact on the environment, the effect of which could be evaluated as a reward function, and the model is improved based on the feedback effect (FSB 2017).

the advantage of being easily communicated. Decision trees are commonly used in medicine for diagnosing and treating.

In ML, decision tree models are both useful for classification and regression modeling. However, they are not often directly used in practice. Two successful extensions of decision trees that have demonstrated high predictive performance are among the most common machine learning tools used in practice: random forest and gradient boosting decision trees. After introducing the basics of decision tree models, these two methods will be presented.

Decision Trees[13]

A decision tree adopts a hierarchical structure like a flowchart that starts from a root node, progresses to lower nodes through possible states or decisions (represented as a branch), and ends at the terminal node that shows the consequence of the entire branch. Decision trees have the advantage of being used for both regression and classification models.

It is easier to demonstrate how a decision tree model works with a chart (see Figure 2). Consider a *regression tree* whose objective is to estimate the LGD using a dataset that contains two features of loans, including the loan-to-value (LTV) ratio and the DTI ratio, and the LGD for each loan. An estimated regression tree is displayed in Figure 2. The tree splits the sample space into five regions. The first split is based on the LTV with a threshold of 0.80. If the LTV is less than 0.8, the DTI threshold is 0.50, but if the LTV is more than 0.8, then the DTI threshold is 0.36. Moreover, for a loan with an LTV<0.8 and a DTI>0.5, the tree further splits the LTV, with a threshold of 0.3. The numbers shown at the end of each branch of the tree are an average LGD of all the loans contained in each region. As a result, the model's prediction for a new loan with an LTV<0.30 and an DTI>0.50 is 4 percent. The tree's depth is 3, which measures the number of splits in the longest branch.

In its simplest form, a tree is constructed by making two decisions at each node until a given criterion is met[14]: (1) which feature to make the split upon; and (2) the associated threshold value for splitting. The objective is to minimize the mean square error of the LGD in the region, that is, find a region with average LGD (the predicted value) that is closest to LGD values of loans belonging to that region. Therefore, a decision tree finds loans that are similar to each other based on splitting the feature space.

A classification tree, that is, a tree to model PD, is developed in a similar way except that the assigned value at the terminal node is the fraction of observed defaults among loans in the region. The decision to split at each node is made to maximize a metric called *purity*, that is, a split to ensure a group of loans with similar default rates fall in the same region[15].

Decision trees should have a stopping threshold to control the size of the tree (for example, the maximum depth of the tree, the minimum number of observations remaining in the final node). Furthermore, trees are *pruned* to avoid overfitting.
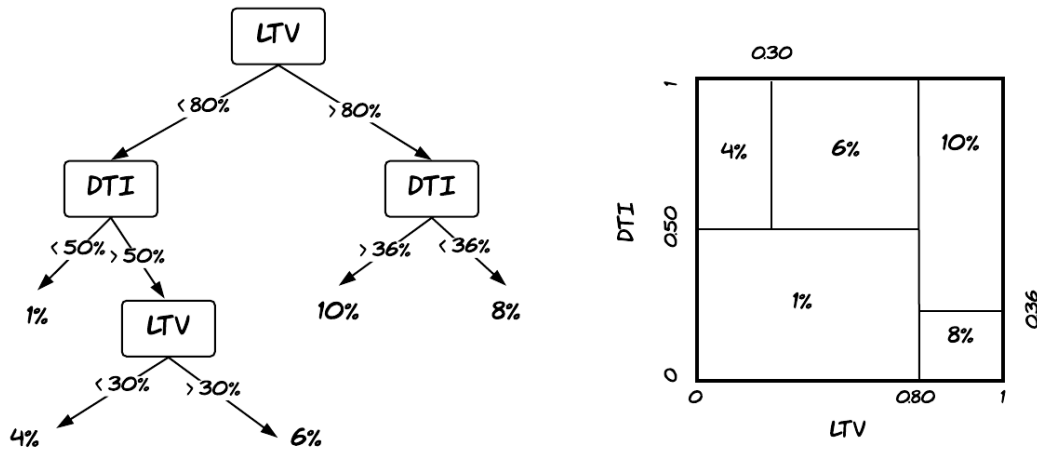
---

[13] Breiman et al. (1984).

[14] Extended versions of trees allow for making multiple splits at each node.

[15] Common measures of impurity are misclassification, Gini, and the Entropy.

Figure 2. Left: An Illustrative Decision Tree Model for Estimating LGD. Right: Partitioning of the Features Space Implied by the Estimated Tree.

Note: Labeled numbers represent average LGD values corresponding to the region. LTV is the Loan-To-Value ratio and DTI is the Debt-To-Income ratio.

Two main advantages of decision-tree models are:

- They are easy to interpret and illustrate for small trees. As a result, they can be easily used to offer clear guidelines in decision making, which is especially important in training credit raters. The guidelines could be further examined against intuition and expert opinion.

- In addition to single-output cases, decision trees can handle multi-output problems. For example, in addition to the LGD, the time to resolution of loans reflecting liquidity risk could be an important factor in risk management.

However, decision trees have certain disadvantages that don't make them the prime method to use, especially when dealing with large data, as is the practice in FinTech lending. Three main practical disadvantages of decision trees are:

- They can create complex trees with unintuitive criteria that are prone to overfitting and therefore do not generalize well to make robust predictions.

- An optimal decision tree is subject to the curse of dimensionality, and in practice they use locally optimal solutions (so-called "greedy" algorithms), which cannot guarantee globally optimal trees. Consequently, they can be unstable and noise sensitive where small variations in the data could sometimes lead to a completely different tree.

- Decision trees are biased when some classes dominate in the sample. This is particularly important in credit analysis, especially when evaluating the credit of underserved populations who don't have a long history or a reliable track record that reflects their creditworthiness.

Motivated by these limitations, extensions of decision trees have been developed that mitigate some of these constraints. These extensions take advantage of *ensemble learning*

methods. Ensemble methods combine several machine learning techniques (for example, trees) into one predictive model in order to decrease variance or bias, or to improve predictions. Random forest and gradient boosting decision trees are two prominent ensemble methods that are based on decision trees.

Random forest[16]
Random forest, or forest of randomized trees, is a technique that combines many decision trees based on two principles: (1) *bagging* technique; and (2) *decorrelating* underlying trees[17]. The first principle, *bagging*, is a general-purpose procedure in which the base model (a decision tree) is estimated on many randomly drawn subsamples with replacement from the train dataset where the subsamples have the same size as the train dataset (also called "bootstrapped" sampling). The predicted values by estimated trees are then averaged to yield the final prediction. In classification, a majority vote is taken to determine the final prediction[18].

Based on the second principle, *decorrelation*, only a subset of features is considered at each node when the decision to split is made. For instance, if there are 16 features in the model, at each node only four are considered to find the optimal split. Furthermore, these four features are randomly selected. This way, each variable has the chance to be used in modeling the outcome variable without having to compete with a dominant feature. As a result, the underlying trees are "decorrelated." A rule of thumb for choosing the size of the subset of features to select at each node is the square root of the number of features. For example, if the number of features (risk drivers) is p, then the number of features to be included at each node is about $\sqrt{p}$. Therefore, if there are 105 features, the random forest method randomly chooses $10 \left(\approx \sqrt{105} \right)$ features and uses them at each node to find the optimal feature to make the split upon.

Figure 3 illustrates different steps of random forest. First, the entire sample is randomly split into train and test datasets. At the bagging stage, random subsamples of the train dataset are generated, and each subsample is fed into estimating the trees. To develop the trees, include **m** chosen features. For example, if a new loan has characteristics displayed in the figure by blue nodes, then its LGD is predicted according to the average values of predicted values by all trees of all subsamples.

The random forest model is widely used in many real-world applications, including banking, medical studies, stock markets, and e-commerce. The two most prominent applications of random forest in banking are (1) identifying loyal customers for lending—in line with this paper's discussion of credit assessment—and (2) detecting fraudulent activities. Random forest is a powerful technique in identifying outliers, which is why they are used in applications such as fraud detection, identifying spam emails, rare diseases, etcetera.

---

[16] Breiman (1996, 2001).
[17] For a more technical analysis of random forests, see Baiu (2012).
[18] The gain in lowering the variance of estimation based on bagging is based on the fact that the variance of the mean of n independent random variables with variance σ^2 is given by σ^2/n, hence a lower variance.

An extension of random forest is extremely randomized trees proposed by Geurts, Ernst and Wehenkel (2006), which makes two important modifications to the random forest algorithm. First, the criterion of splitting is not based on an optimized split, that is, one that minimizes prediction errors of implied partitions (Figure 2), but thresholds are randomly drawn for each candidate feature and the best of these randomly generated thresholds is picked as the splitting rule. Second, extremely randomized trees do not use bootstrapped samples but the train dataset. These two modifications lower the time of processing to a great extent while offering comparable performance to a random forest. Extremely randomized trees are applied in the same areas that random forests are. A well-known use in medicine is medical image analysis[19].

Figure 3. The Random Forest Model
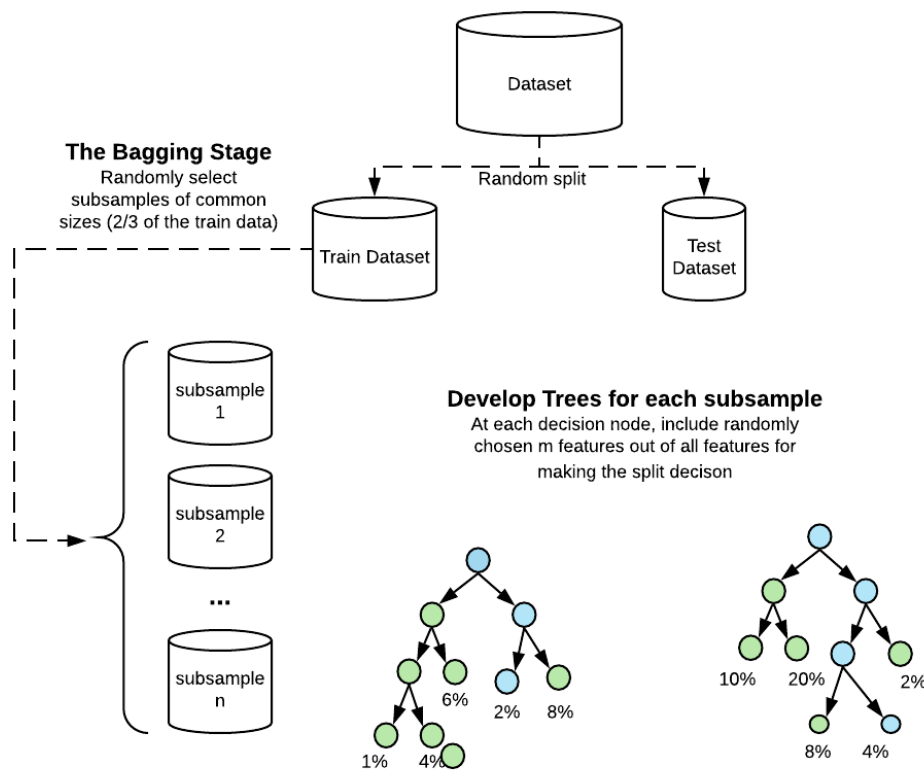
The random forest model improves upon the decision trees and has the following two advantages:

- Relative to decision trees, they are less subject to overfitting, have superior generalization properties, and can make more robust predictions, which is why they have attracted practical interest in various fields.

- Unlike decision trees, a dominant feature does not lead to the results in the random forest model due to the decorrelation stage, which provides the chance for various features to

---

[19] See Criminisi & Shotton (2013).

contribute to the final prediction value. For the same reason, the random forest model can easily handle a large set of correlated features.

At the same time, the random forest model yields complicated rules, which can be hard to interpret and communicate, especially when the feature space and the number of generated trees is large. One way to finesse this problem is to rank features based on their contribution to the final prediction, a technique called "feature importance" in machine learning. For regression models, feature importance is based on the sum of reductions in residual sum of squares caused by nodes in which a specific feature was the basis for splitting. Similarly, in the context of classification, feature importance adds up to the reduction in the Gini impurity index[20] at each split resulting from a specific feature. The relative ranking of features allows for identifying features with the highest role in making predictions by the estimated model.

Feature importance could play a big role in assessing the results of credit analysis based on a random forest model against intuition. If the identified features are widely different from the expected set of features, the model or the data should be evaluated more closely.

Gradient-Boosting Decision Tree (GBDT)
GBDT is another successful tree-based ensemble machine learning technique that, like random forest, has attracted a lot of interest in real-life applications. A well-known example is finding high-quality user-generated content in social media through Yahoo[21]. Another example is applying GBDT to learn to rank for information retrieval, which is a key component in automated data analysis[22].

Instead of building and combining parallel and independent trees as in random forest, GBDT builds a series of trees, each trained based on the mistakes of the previous tree[23]. Specifically, at the first step, a simple tree is estimated on the entire train dataset. The second tree is estimated using prediction errors of the first tree as the output variable so as to decrease the prediction error of the first tree. Other trees are built sequentially in the same manner. The final prediction is the weighted sum of predictions of all trees. The weighting is governed by a key parameter specific to GBDT called the *learning rate*.

It has been empirically shown that using a small learning rate (for example, less than 0.1) with a large number of trees dramatically improves the model's out-of-sample predictive power compared to a learning rate equaling one with a smaller number of trees. Clearly, GBDT with a small learning rate comes at the cost of higher computational time because of the need for more iteration to achieve a well-functioning GBDT.

The baseline trees in GBDT are generally shallow, with only a few branches, and are therefore considered weak learners, that is, each tree only adds a small amount to the

---

[20] The Gini impurity is computed by summing the probability of an item with a label being chosen times the probability of a mistake in categorizing that item.
[21] See Agichtein et al. (2008).
[22] See Liu (2009). Liu defines learning to rank for information retrieval as "a task to automatically construct a ranking model using training data, such that the model can sort new objects according to their degrees of relevance, preference, or importance."
[23] Friedman (2001).

predictive power. However, many trees are added sequentially, with the goal of minimizing the error made by the previous tree. In calculating the final prediction, GBDT weights the predictions of all trees by the degree of correction that each tree can make to errors. This way, GBDT aggregates the predictive power of many weak learners to produce a strong predictor. In general, GBDT is considered to be superior to random forest, which performs better than basic trees.

Because the number of trees is usually very large, it is not feasible to display in a figure the underlying steps that an estimated GBDT takes in making predictions. However, like random forest, it is also possible to rank contributions of features in predicting the outcome, that is, generate feature importance. Another tool that summarizes the findings of an estimated GBDT is the partial dependence plot (PDP). PDPs show the dependence between the response variable and a feature by marginalizing over the values of all other features. PDPs can be interpreted as the expected effect on the response variable as a function of a feature while keeping other features fixed at their average.

As an extension of GBDT motivated by the bagging technique, Friedman (2002) proposes *stochastic gradient boosting* that improves upon GBDT by using a random subsample of the training data (like random forest) rather than the entire train data when estimating a tree at each iteration. A fraction between 0.5 and 0.8 of training data leads to best performance. Furthermore, stochastic gradient boosting decreases the computational time because it uses a smaller sample to estimate the trees at iteration.

Support Vector Machines (SVMs)
An SVM is a machine learning technique that is based on an intuitive approach that finds the best separation between observations that belong to different classes[24]. While SVMs were developed primarily to solve classification problems, such as face detection, handwriting recognition, image classification, and bioinformatics, their extension allows for modeling regression problems as well.

Consider the problem of predicting the default outcome of loans based on two features: the DTI and LTV ratios. Figure 4 displays a sample SVM estimated for this model. The objective of an SVM is to find the best separating line[25] that splits the feature space based on default outcome. Unlike trees that make splits based on a single feature and therefore result in box-shape regions, SVMs find the border of separation based on a function of all of the features. To find the best separating line, the SVM tries to maximize the minimum distance between the line and closest points in each class, which is known as the *margin*. As shown in Figure 4, the classes are not always completely separable. In this case, the SVM compromises some points (hence misclassifying those) and finds the separating line to generate the most homogenous observations in each region.

---

[24] Cortes et al. (1995) and Ben-hur et al. (2001).
[25] In higher dimensions, the separating line is called "separating hyperplane."

The decision rule implied by a simple SVM, as displayed in Figure 4, is quite intuitive. For instance, "if 0.5×LTV+0.8×DTI is greater than 3, the loan is labeled as risky otherwise safe."

Figure 4. An SVM Model for Predicting Default Outcome Based on Debt-To-Income (DTI) and Loan-To-Value (LTV) Ratios.
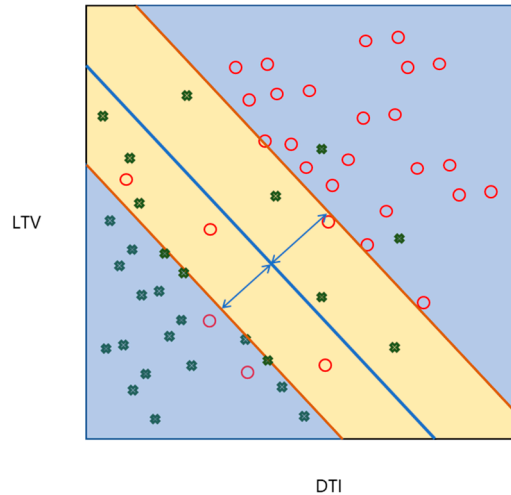
Note: Red circles represent "no default" and green crosses represent "default." The thick blue line is the separating line estimated by SVM; red lines are support vectors distanced from the blue line by the margin.

The functional form that governs the shape of the separating line is linear. In this case, the SVM is based on a linear *kernel*. Natural extensions of the SVM when the data are not linearly separable is to use nonlinear kernels, which could therefore yield nonlinear decision boundaries. In effect, a nonlinear kernel is a transformation of the feature space, which could generate a sharper separation of outcome classes. Common nonlinear kernels used with an SVM are polynomials and radial basis function (RBF) kernels. Examples are displayed in Figure 9. It is also possible to construct a customized kernel.

SVMs offer three main advantages:
- Because they use a subset of the training dataset to evaluate the separating line and the support vectors, SVMs are memory efficient. Likewise, they scale well for high-dimensional problems and they are less prone to overfitting.

- The nonlinear kernel gives a high flexibility to SVMs, especially when the shape of the kernel is motivated by business knowledge and intuition about the problem. However, the performance of SVMs is sensitive to the choice of kernels.

- SVMs even work well with unstructured and semi-structured data like text and images.

However, their main disadvantage is:
- Estimated SVMs are difficult to interpret. For instance, it is hard to understand what variable weights represent and how individual features impact the outcome variable. In

this sense, they are more black-box models. Moreover, they do not directly provide probability estimates, which is particularly useful in estimating the probability of default[26].

Figure 5. SVM With Linear and Nonlinear Kernels

| Linear kernel | Polynomial kernel | RBF kernel |
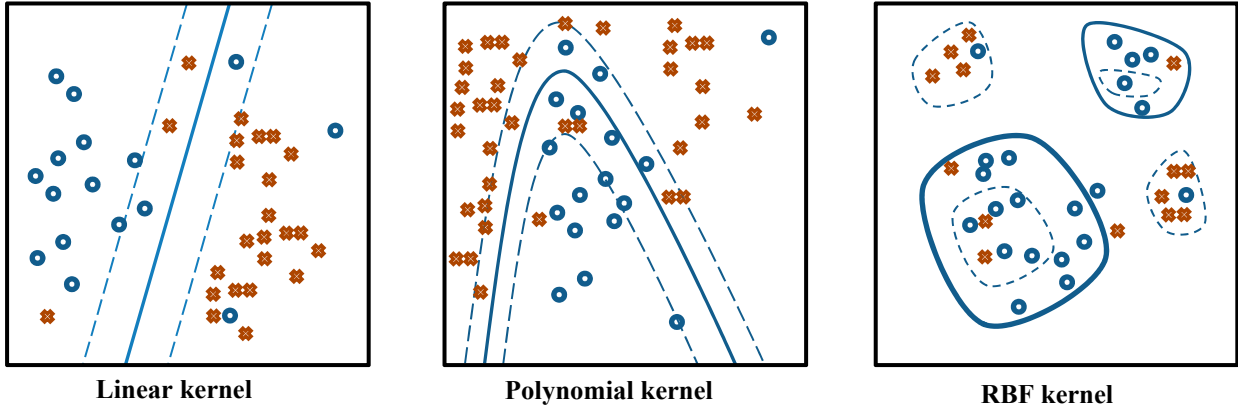
Neural Networks

NNs are another class of powerful predictive ML models that have become widely used, for example in the areas of computer vision and natural language processing. NNs were primarily motivated by the human brain (hence their name[27]). An example neural network structure is shown in Figure 6. NNs are built to learn the pattern between features and output through several inner layers. At the first layer, features (denoted by X_1 to X_4 in Figure 6) are used to evaluate the value of nodes Z_1 to Z_6, which are then used as input for calculating nodes of the second layer. The calculation is based on a function on a weighted sum of inputs where the function's output range is between 0 and 1.[28] A common functional form is the sigmoid $\sigma(x) = \frac{1}{1+e^{-x}}$. For example,

$$Z_1 = \frac{1}{1+e^{-(\omega_1 X_1 + \cdots + \omega_4 X_4 + b)}},$$

where ω_1 to ω_4 determine the contribution of each feature to the node and the constant b is intended to capture the *bias*. The above form resembles the logistic regression and therefore, in their basic form, NNs can be seen as comprising many logistic regressions, the outputs of one layer used as inputs in lower layers.

To evaluate each Z in layer 1, five coefficients (four weights and a bias) are needed. Adding up for five nodes, this means layer 1 has 25 parameters. Suppose layer n was the only other hidden layer of NN in Figure 6. To estimate the three nodes in this layer, 18 more parameters should be estimated. In sum, the two-layer NN has 40 parameters to estimate (compare to five coefficients for an ordinary linear regression). It is immediately clear that a deeper NN

---

[26] It is possible to use workarounds such as fivefold cross-validation, but this approach is usually time-consuming when the sample size is large.

[27] This class of models is sometimes called "artificial neural networks," particularly in the field of computational neurobiology, where the term "neural network" is reserved for the study of the human brain.

[28] The function is commonly referred to as *activation function*, which acts like a classification model, which combined with a threshold, leads to activate 1 or inactive 0 outcome from a node. Other activation functions include hyperbolic tangent, softmax, and rectifier functions.

would have many parameters, thereby requiring a large data to train, a more complicated optimization algorithm, and high computational power.

Extended versions of NNs, famously called *deep learning,* are one of the high-impact areas of ML models that have been successfully deployed in numerous fields due to higher efficiency and reasonable processing time[29]. Examples include computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, and bioinformatics, where they have produced results comparable to and in some cases superior to human experts[30].

The main advantage of NNs is their flexibility and ability to recognize complicated patterns in the data, which is particularly useful when the size of the data is large. As the sample size becomes larger, performance gains of traditional models remain low, whereas NNs are able to capture deeper connections between features and the output. However, those patterns are not understandable by human beings, making NNs black-box models. In effect, NNs learn by memorizing patterns in the data through the estimation of a large set of parameters. In addition, NNs use nonlinear structure via the activation function, which increases the model's flexibility.

Figure 6. A Neural Network with Four Input Variables (Features), Two Outcomes, and n **Hidden Layers.**
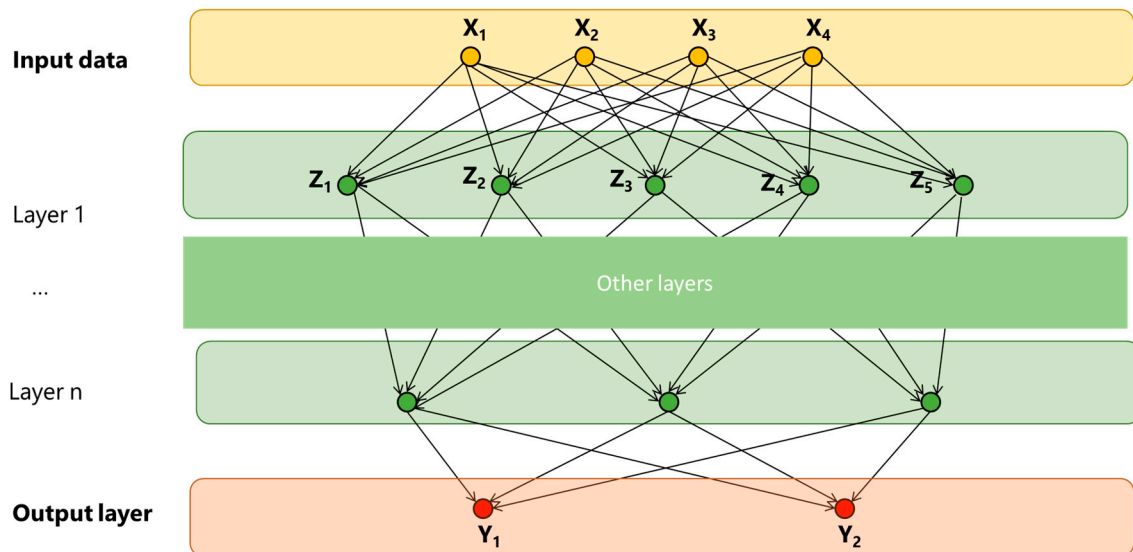
A challenging issue in using NNs is the way business knowledge should be incorporated in the model to improve the model's performance. This could only be done through modifying the feature's input to the model, but the structure of deep NNs could not be modified to reflect non-data information. This constraint of NNs has made it less attractive to be used as a standalone technique for lending purposes, especially when there is the risk of data contamination.

C. How Does Machine Learning Differ from Econometrics?

With this background, I turn to discuss some of the main differences between mainstream econometrics and ML approaches. While a detailed examination of the commonalities and differences among these two approaches is beyond the scope of this paper, the following discussion focuses on their most prominent distinctions in data analysis.

The literature in ML and econometrics significantly overlap, as the main objective of both is to learn from data. Both estimate models that explain the relationship between features (explanatory variables) and the output (response) variable that are intended to guide decision making. However, one broad way of making a distinction between the two approaches is that the main focus of ML models is to make the best out-of-sample prediction of the outcome variable without necessarily trying to explain the structure of the underlying relationship[31], whereas mainstream econometrics aims at finding casual relationships between observed features and the outcome variable[32].

The main challenge in data analysis that econometric studies attempt to address is how to deal with selection bias, which occurs when the relationship between features and the outcome variable differs between the observed data and the intended causal effect of a decision variable.[33] The famous result of the selection bias problem is that correlation does not imply causation, and therefore finding a high correspondence between the features and the outcome variable does not necessarily inform how the outcome will respond to a decision or a policy action in the presence of selection bias. The identification of the causal effect is guided by theory and restricting assumptions to infer how features influence the outcome variable.

However, econometric studies rarely look for maximizing the out-of-sample forecasting power of the model, which is the central focus of ML. The problem of selection bias relates to the ML modeling's assumption of generalizability of the observed data for making predictions for the unobserved sample. This means that the prediction problem should be

---

[31] A recent literature in ML and artificial intelligence has tried to make ML models more interpretable. See Gilpin et al. (2018) for a review. Moreover, econometric tools such as time-series analysis are often used for forecasting purposes. Nonetheless, these methods are built to have some aspect of interpretability of the underlying parameters.

[32] Both literatures share many common tools; for example, nonparametric models such as kernel regressions appear in econometrics, and some ML models such as linear discriminant analysis, naïve Bayes, and even a simple neural network could be considered parametric models. This section broadly highlights the main differences in the mainstream practice of ML and econometrics that are relevant in credit risk assessment.

[33] See, for example, standard econometrics textbooks like Angrist and Pischke (2008) and Wooldridge (2015) for elaboration.

carefully defined and the relevance of the sample for making prediction should be verified. For instance, in the case of assessing the credit risk, the sample should be representative of the borrowers who will be assessed using the model. For example, a model that is trained based on data that only contain large companies may not capture risk characteristics of loans to small- and medium-size enterprises because of underlying differences between the two loan classes.

Unlike ML, econometric studies pay close attention to the statistical significance of an estimated parameter. As long as a variable can increase the predictive performance of the model, it will be included in the ML modeling. In a similar vein, in ML model parameters do not often have useful interpretations to generate business insight. For instance, the number of trees used in random forest or GBDT, the margin parameter in SVM, and the depth and nodes of neural networks govern each model's structure and flexibility to learn from data, but do not add any insight in terms of credit risk assessment. Furthermore, while it is possible to draw decision rules from an estimated ML model, because of their high complexity, these rules could often only be applied by machines rather than by human agents. While this is useful for automating credit decisions, it has the shortcoming of making it difficult to evaluate the soundness of decisions against intuition and business knowledge.

IV. STRENGTHS AND WEAKNESSES OF ML-BASED LENDING FOR FINANCIAL INCLUSION

Building on previous sections, this section discusses potential advantages and shortcomings of using ML for credit rating and how ML-based lending could influence financial inclusion focusing on the credit aspect of financial inclusion. Because undue credit growth could lead to financial instability, the paper's focus is on promoting prudent lending in the economy by FinTech. Admittedly, there are other important enabling factors for prudent lending such as sound regulation and supervisory regimes, depth of the financial system, and high financial literacy of borrowers that are beyond the scope of this paper.

A. Strengths of ML-Based Lending

FinTech companies aim at changing the way traditional lending functions by automating the steps of credit assessment and lending decisions. This is carried out by training ML models through information gathered from a wide range of sources.

ML can make assessing credit risk of small borrowers feasible and economical. Banks often refrain from assessing the credit risk of a small borrower because the small size of the loan and potentially high risk of the loan do not justify the cost of employing a professional examiner and meeting underwriting standards. The sunk cost of initiating a credit assessment is a major factor in financial exclusion of many households and small businesses (*World Economic Forum 2016*).

However, automated credit assessment allows FinTech companies to conduct credit scoring of small borrowers much more frequently than traditional lending. By automating the credit rating process, FinTech credit companies could better assess the creditworthiness of borrowers through making small amounts of loans at a high frequency and monitoring the repayment behavior of the borrower. This way, FinTech credit companies can assess

eligibility of borrowers for receiving loans rather than rationing those borrowers who lack sufficient standard data such as financial reports used in traditional credit risk assessment. [34]

Another way that nonbank lenders could obtain massive information about potential borrowers is the experience of the so-called "BigTech" lending companies.[35] BigTech companies have the advantage of accessing information about customers' business activity as a byproduct of offering e-commerce, payment and telecommunication platform services (*World Economic Forum 2015*). Big tech firms' marketplace allows for observing the sales trend and cash flows of businesses active in their e-commerce space, thereby extremely reducing the cost of collecting information. Using ML analysis, they could therefore process the creditworthiness of borrowers without requiring professional financial reports, which is often cited as a big burden for small borrowers. Furthermore, they can make cashflow-based loans without requiring collateral, another major hindrance for small companies trying to obtain credit from banks (BIS 2019).

Similarly, big tech companies that offer payment services to consumers can leverage transaction data to acquire more accurate information about consumption behavior and income stream to assess creditworthiness of borrowers, therefore removing the need for an established credit registry to collect consumer information. This way, big tech lenders could provide financing to consumers who are unable to take personal loans from banks and improve financial inclusion.

For an empirical example, Schweitzer and Barkley (2017) study the Federal Reserve's Small Business Survey and find that business borrowers who received funding from online platforms share similar characteristics with those who were denied credit, suggesting FinTech lending contributed to financial inclusion of these small borrowers.[36]

ML can harden soft information
ML models are well known for performing particularly well in the presence of hard information (Liberti and Peterson 2018). According to them, hard information is characterized by two important properties: (1) it has a numeric value for data entries that measures a feature in a cardinal way; and (2) it is independent of the data collection process. Variables such as income, debt, number of employees, and sales are some examples of hard information that could be used in estimating the debt servicing capacity of the borrower. Similarly, among other Cs of credit rating, capital structure, coverage, and conditions are generally based on hard information. Price variables used for evaluating these attributes are hard information as long as they are inferred from market data.

However, a major challenge in evaluating asset values of small borrowers is that there is no liquid market for discovering the price. The common practice among traditional lenders in

---

[34] According to data by TransUnion, FinTech credit accounted for 38 percent of total personal loans in the US in 2018, rising from only 5 percent in 2013 (see CNBC article on Feb 21, 2019).
[35] The term "BigTech" refers to the direct provision of financial services or of products very similar to financial products by technology companies (BIS 2019).
[36] Using data of consumer credit in Germany, de Roure et al (2016) find that the risk-adjusted interest rates on P2P loans are lower than those on bank loans.

this case is to rely on a specific pricing model for the asset, which therefore contains many judgmental factors by the assessor. Similarly, the borrower's character is evaluated based on the rater's subjective assessment and may not easily generalize for other cases. Furthermore, information that a lender obtains through maintaining a long-term relationship is borrower-specific and judgmental, and therefore soft information.

Nevertheless, the challenge with processing soft information by ML-based analysis is a data limitation issue. Once sufficient data that are not predominantly noisy are collected, classification ML models could be used to turn soft information into hard information. Thanks to the high capacity of ML models to include large and diverse data, ML-based lenders can turn almost anything into data, ranging from social media activities to the physical location of applicants' activities. ML models offer the possibility of enhanced preprocessing and incorporating unstructured data in the analysis, which facilitates hardening soft information of data collected from a wide range of sources.

As a result, ML models could enhance pricing of assets pledged by small borrowers by finding common patterns in the nontraditional data—they could build reliable and powerful pricing models[37]. Similarly, social media information could be used to sharpen evaluation of the borrower's character. For instance, Jagatiani and Lemieux (2017) show that using nontraditional data resulted in more informative credit scoring by FinTech lending firms in the US in comparison to the traditional FICO score system. Likewise, Berg et al. (2018) show that easily accessible information from digital footprints could lead to superior performance of credit rating relative to credit bureau scores.

ML can better capture nonlinearities
By searching for relationships in small partitioning of the sample (such as tree-based models, discussed in the previous section), ML models can capture local relationships between risk indicators and credit risk outcomes that traditional models or human agents fail to detect. As long as extra information collected and used by ML models is informative about borrowers, ML-based models can outperform traditional methods by capturing higher dimensions of nonlinearity. For example, through effective partitioning of borrower features, ML may find credit indications for those borrowers who have been identified as poor credit based on traditional indicators and denied credit. This way, by better capturing nonlinearities in the relationship between the credit risk outcome and features, ML is capable of better assessing the borrowers, particularly those with poor credit history, and avoiding being driven by the majority. Specifically, high-quality borrowers, due to their favorable historical record, dominate the sample and tend to drive the results in their favor if the model does not distinguish between different sections of the sample (for example, logistic regression). However, methods like random forest can focus on a particular class of borrowers by partitioning the sample and detecting possibly strong relationships for those borrowers. Therefore, while borrowers with a poor credit history based on traditional indicators of creditworthiness may be denied credit by traditional lenders, FinTech lenders may find them eligible to borrow funds.

---

[37] See, for instance, Gu, Kelly, and Xiu (2018).

ML could mitigate information asymmetry

Another layer of complication arises in lending due to information asymmetry between the lender and the borrower, which has created a large literature in corporate finance. Moral hazard and adverse selection problems could arise when the lender cannot reliably obtain information about the borrower, and the borrower has an incentive to misreport in order to obtain better loan terms, but taking this possibility into account, the lender would ration credit to avoid making losses. Such rationing behavior becomes extreme, as information asymmetry is more intensified (Tirole 2010).

ML could potentially mitigate financial exclusion due to information asymmetry by identifying indicators of borrower riskiness through analyzing larger data and identifying differences among borrowers. This could be accomplished by detecting nonlinear relations between features and credit risk that are otherwise unidentified by traditional approaches, leading to adverse selection and moral hazard problems. Modern technology allows for more enhanced monitoring of borrowers by screening the borrower more frequently and effectively. For instance, text, image, and video data could be collected on a company's business operations and could be used to assess compliance with loan covenants. Such enhanced monitoring could help to mitigate moral hazard problems.

While ML could in principle lead to better evaluation of the credit risk of borrowers, it could have heterogeneous impact on borrowers leading to the exclusion of a subset of borrowers, because the lender can make more informed credit decisions. Such financial exclusion is, however, consistent with prudent lending and sound risk management practices and could result in higher profitability of lenders and potentially superior financial stability.

B. Weaknesses of ML-Based Lending

ML-based lending bears risks of financial exclusion

ML models are trained using available data that may not necessarily be representative of all classes of borrowers that the creditor considers lending to—a situation that violates the principle of generalizability. As a result, the lack of sufficient relevant data for some classes would impose restrictions on the conclusions by ML analysis and could lead to redlining applicants belonging to a particular group. A case in point is a sample with a class of borrowers that have been historically denied credit for nonbusiness reasons (for example, gender, race, religion, ethnicity, residence in certain areas, etcetera). If the raw sample is used for making predictions regarding default behavior of underserved borrowers, the results will be contaminated by the effect of traditionally unfair financial exclusion of those borrowers, and the model will further predict poor credit rating. In such cases, features including gender, race, and religion that could cause bias should be excluded from the model. Such variables could, however, be used to assess the existence of bias by a scoring model and avoid discrimination.[38]

---

[38] In this regard, a line of research studies discrimination-aware data mining. See Berendt & Preibusch (2014), Lepri et al. (2017) and Žliobaitė (2017).

ML-based credit rating could cause consumer protection, ethical, and data privacy issues

The use of information gathered from large pools of data (including social network and internet of things) not only raises an issue around the relevance of the information (the noise) but further compounds concerns on consumer protection, as exclusion could stem from using personal data to exclude consumers on the basis of unexplainable computer decisions[39]. Therefore, in addition to the risk of unjustified exclusion, there is a risk to consumer protection because of the opaqueness of ML-based credit rating. As ML-based credit rating is often based on a large set of risk drivers[40], it may be hard to detect and prove the dominance of one factor in the final credit rating outcome, which creates space for unethical decisions and impedes supervision.

While this is a common problem between ML and traditional credit rating, it could be more difficult to detect with a large set of risk indicators. One way to mitigate such problems is to ensure that the set of features does not include discriminatory variables, and hence there is no role for such variables to drive the results. Supervisors could enforce this requirement by monitoring the set of input variables in the credit risk modeling process. Moreover, as discussed in the previous section, ML models could provide top variables that contribute most to the credit score, that is, by using "feature importance" of each algorithm. Lenders should actively monitor most significant drivers of credit rating of borrowers and assess against their business insight. Feature selection techniques such as LASSO could be used to reduce the number of attributes that drive the credit score, which is another way to make analysis of data with many features more tractable and thereby easier to monitor.

ML may not address structural changes

A similar challenge arises when a large change in the environment renders historical data irrelevant for assessing new customers, that is, analysis of the historical sample does not generalize for new applicants because of significant structural changes (for example, a change in financial development, macroeconomic policies, and industrial changes)[41]. If the modeler fails to identify such dramatic changes, the trained model's risk assessment could have costly implications for the creditors. Of course, if there are historical observations of similar structural changes, ML could capture changes in the environment.

One shortcoming of ML models is that the underlying algorithm cannot be easily modified to reflect tacit information such as business knowledge, intuition, and anticipation of future events not realized in the data. Specifically, as discussed in the previous section, ensemble tree-based models such as random forest grow many trees and randomly select features, and therefore their underlying algorithm is too inflexible to incorporate tacit knowledge. Likewise, deep NNs have many parameters without a clear link to a particular piece of information and thus offer little modeling flexibility in this regard. SVMs have some limited flexibility in the choice of kernels, but it is not exactly clear how kernel choices could be

---

[39] For example, Fuster et al. (2018) propose a cross-category measure of disparity and find that ML models could worsen disparity relative to the logit model according to this measure in the US mortgage sample.
[40] According to Claessens et al. (2018), "the website of one Indian P2P platform claims that its credit assessment involves a review of more than 1,000 data points per borrower."
[41] See, for instance, Rajan, Seru, and Vig (2010).

modified to reflect complex knowledge. By contrast, in a Bayesian framework, it is sometimes possible to build intuition-type information into priors or the shape of the likelihood function.

Faking indicators by borrowers could damage ML-based scoring
If borrowers realize that an indicator is important for their credit rating, they are motivated to artificially change its value if they can. For example, the number of social media connections could be indicative of a high character score. However, considering it does not reflect the strength of relationships, applicants could expand the size of their network to achieve better credit rating. In that case, the information content of the variable changes over time and a model that is trained based on old data would yield erroneous credit scoring for new borrowers.

V. Concluding Remarks

I conclude by reviewing the main points of the paper and a brief discussion of the potential for FinTech credit in emerging economies with poor financial development. Achieving widespread financial inclusion is a major step in achieving sustainable development goals in many countries, and FinTech credit appears as a promising solution and a potential leapfrog for countries with low financial inclusion.

One of the core strengths of FinTech credit is the use of machine learning techniques and big data analytics to promote credit scoring. This paper reviews the underlying challenges in assessing credit risk of particularly small borrowers and discusses the most prominent ML techniques applied in assessing credit risk of borrowers for a nontechnical audience. It provides sufficient background to analyze general strengths and weaknesses of using FinTech for increasing financial inclusion.

Among the strengths of ML in promoting financial inclusion, it could particularly facilitate low-cost automated evaluation of small borrowers that would be otherwise left out of the traditional credit market. By turning soft information to hard information, ML allows for incorporating a wide range of information from various sources and structures. Moreover, ML can capture nonlinearities in the relationship between risk drivers and credit risk outcomes substantially beyond traditional models. These strengths could enable ML to mitigate the information asymmetry problem that is at the heart of lending, particularly to small borrowers with a short history of formal financial reports.

However, ML has certain weaknesses that should be taken into consideration when applying for credit risk assessment. Heavy reliance on learning from data, particularly in a context where the size of the sample is considerably larger than traditional ways of credit scoring, could result in noisy information playing a role in driving results of credit analysis and leading to financial exclusion of creditworthy applicants. The bias in the sample should be identified and avoided by analysts as much as possible to avoid digital financial exclusion. Along the same lines, ML may not capture structural changes in a timely manner, because

arrival of informative data may be slow due to the lengthy process of default observation.[42] This could negatively impact FinTech lenders, who rely on ML to assess borrowers without evaluating the relevance of data used for training the model for new applicants. Furthermore, borrowers may realize and counterfeit certain indicators that drive their credit score, thereby decreasing the relevance of those features for new applicants. Finally, it should be noted that ML is exposed to some of the key concerns in econometrics, most importantly the endogeneity and selection bias problem. The analyst should check the sample to ensure proper treatment of these issues and avoid superfluous results. Nonetheless, proper choice of risk drivers makes these issues less of a concern for credit risk assessment and default outcome prediction.

Taking these strengths and weaknesses into consideration, emerging market economies (EMEs) that are challenged by low financial inclusion could highly benefit from FinTech lending. ML-based credit rating could have important consequences for the credit market in EMEs. By enhancing credit analysis, FinTech lending can reduce the time of credit decision and offer lower loan rates to existing borrowers. Lengthy credit decisions and high loan rates are among the main financial constraints in EMEs that confine the pace of the private sector growth. Banks and FinTech lenders benefit by appropriately pricing the risk, thereby lowering debt servicing costs for borrowers while improving their asset quality.

FinTech credit could increase financial inclusion by lending to underserved population relying on enhanced credit rating given that credit bureaus are either nonexistent or poorly developed in most EMEs. The asset quality of these borrowers will likely be lower than traditional borrowers, which implies the lenders take on more risk and should therefore offset the risk by holding sufficient buffer reserves and capital. ML evaluation of credit risk should allow for measuring the amount of buffer.

Nevertheless, EMEs may face some challenges in reaping these benefits while ensuring that the development of FinTech credit does not expose systemic risks to the financial system. First, given the central role of data in ML-based credit analysis, it should be legally and technologically possible to gather digitalized data reliably from various sources and avoid noisy and biased data as much as possible. The lower the noise of the data and the less need for data cleaning, the more effective the results of credit analysis. As a complement for high-quality data availability, cybersecurity measures should be in place because of the sensitivity of credit information.

Second, technological infrastructure for implementing big-data-based decision making should be developed and accessible for FinTech credit. Moreover, ML analysts should consistently review ML models for credit rating to avoid potential weaknesses of naïve application of ML. For example, analysts should conduct ongoing monitoring of the environment to identify structural changes, examine indicators that could be faked by borrowers, assess whether the most important drivers of credit risk implied by ML are justified based on business knowledge.

---

[42] Specifically, data of loan performance at least over a full business cycle and a credit cycle is needed to appropriately train the ML model.

REFERENCES

Agichtein, E., C. Castillo, D. Donato, A. Gionis, and G. Mishne (2008). "Finding high-quality content in social media," *Proceedings of the 2008 international conference on web search and data mining,* 183-94). ACM.

Altman, E. I., and E. Hotchkiss (2010). *Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt,* Vol. 289. John Wiley & Sons.

Angrist, J.D., and J.S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press.

Athey, S. (2018). "The impact of machine learning on economics," *The Economics of Artificial Intelligence: An Agenda.* University of Chicago Press.

Athey, S., and G.W. Imbens (2015). "Machine learning methods for estimating heterogeneous causal effects," *1050*(5).

Bank of America (2018). "5 Cs of Credit: What Are Banks Looking For?" available at https://www.bankofamerica.com/smallbusiness/business-financing/learn/5-cs-of-credit/, (accessed December 17, 2018).

Bastos, J.A. (2010). "Forecasting bank loans loss-given-default," *Journal of Banking & Finance*, *34*(10), 2510-17.

Bazarbash, M. (2017). "Loss Given Default of Secured Commercial Loans," August, available at https://sites.google.com/view/majidbazarbash (accessed January 3, 2019).

Bazarbash, M. (2019). "Reliability of Out-of-Sample Forecasting: A Case Study of Loss Given Default of Commercial Loans." Mimeo.

Berkmen, P., K. Beaton, D. Gershenson, J. A. del Granado, K. Ishi, M. Kim, E. Kopp, and M. Rousset (2019). "Fintech in Latin America and the Caribbean: Stocktaking." March, International Monetary Fund Working Paper No. 19/71.

Bengio, Y. (2009). "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, *2*(1), 1-127.

Ben-Hur, A., D. Horn, H.T. Siegelmann, and V. Vapnik (2001). "Support vector clustering," *Journal of machine learning research*, *2*(Dec), 125-37.

Berg, T., V. Burg, A. Gombović, and M. Puri (2018). "On the Rise of FinTechs–Credit Scoring Using Digital Footprints" (No. w24551), National Bureau of Economic Research.

Biau, G. (2012). "Analysis of a random forests model," *Journal of Machine Learning Research*, *13*(April), 1063-95.

BIS (2018). "Sound Practices: Implications of fintech developments for banks and bank supervisors." February, available at https://www.bis.org/bcbs/publ/d431.pdf (accessed January 2, 2019).

BIS (2019). "BigTech and the changing structure of financial intermediation." April, available at https://www.bis.org/publ/work779.htm (accessed May 8, 2019)

Breiman, L. (1996). "Bagging predictors," *Machine learning*, *24*(2), 123-40.

Breiman, L. (2001). "Random forests," *Machine learning*, *45*(1), 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. "Classification and Regression Trees," Boca Raton, FL: CRC Press, 1984.

Berendt, B., & Preibusch, S.(2014). Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence and Law*, *22*(2), 175-209.

Buchak, G., G. Matvos, T. Piskorski, and A. Seru (2018). "Fintech, regulatory arbitrage, and the rise of shadow banks," *Journal of Financial Economics*, *130*(3), 453-83.

CGFS & FSB (2017). "FinTech credit: Market structure, business models and financial stability implications. May, available at https://www.bis.org/publ/cgfs_fsb1.htm (accessed January 2, 2019).

Claessens, S., J. Frost, G. Turner, and F. Zhu (2018). "Fintech credit markets around the world: size, drivers and policy issues," *BIS Quarterly Review,* September.

Cortes, C., and V. Vapnik. (1995). "Support-vector networks," *Machine learning*, *20*(3), 273-97.

CNBC (Feb 21, 2019). "Fintechs help boost US personal loan surge to a record $138 billion." February 21, available at https://www.cnbc.com/2019/02/21/personal-loans-surge-to-a-record-138-billion-in-us-as-fintechs-lead-new-lending-charge.html (accessed March 14, 2019).

Criminisi, A., and J. Shotton (Eds.) (2013). *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media.

Das, S.R. (2018). The Future of FinTech. November, available at https://srdas.github.io/Papers/fintech.pdf.

De Roure, C., Pelizzon, L., & Tasca, P. (2016). How does P2P lending fit into the consumer credit market?. Mimeo.

Friedman, J.H. (2001). "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, 1189-1232.

Friedman, J.H. (2002). "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, *38*(4), 367-78.

FSB (2017a). "Financial Stability Implications from FinTech: Supervisory and Regulatory Issues that Merit Authorities' Attention." June, available at http://www.fsb.org/wp-content/uploads/R270617.pdf (accessed January 2, 2019).

FSB (2017b). "Artificial intelligence and machine learning in financial services." November, available at http://www.fsb.org/wp-content/uploads/P011117.pdf (accessed December 17, 2018).

Fu, K., D. Cheng, Y. Tu, and L. Zhang (2016). ''Credit card fraud detection using convolutional neural networks," *International Conference on Neural Information Processing,* 483-90). October. Springer, Cham.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2018). "Predictably unequal?," *The Effects of Machine Learning on Credit Markets.* November 6.

Fuster, A., M. Plosser, P. Schnabl, and J. Vickery (2018). "The role of technology in mortgage lending" (No. w24500), National Bureau of Economic Research.

Geurts, P., D. Ernst, and L. Wehenkel (2006). "Extremely randomized trees," *Machine learning*, *63*(1), 3-42.

Gilpin, L.H., D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, and L. Kagal (2018). "Explaining Explanations: An Overview of Interpretability of Machine Learning, *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA),* 80-9). October. IEEE.

Gomber, P., R.J. Kauffman, C. Parker, and B.W. Weber (2018). "On the Fintech Revolution: Interpreting the Forces of Innovation, Disruption, and Transformation in Financial Services," *Journal of Management Information Systems*, *35*(1), 220-65.

Gomber, P., J.A. Koch, and M. Siering (2017). "Digital Finance and FinTech: current research and future research directions," *Journal of Business Economics*, *87*(5), 537-80.

Gu, S., B. Kelly, and D. Xiu (2018). "Empirical asset pricing via machine learning" (No. w25398), National Bureau of Economic Research.

Hastie, T., R. Tibshirani, and J. Friedman (2016). *The elements of statistical learning*. New York: Springer.

He, M.D., M.R.B. Leckow, M.V. Haksar, M.T.M. Griffoli, N. Jenkinson, M.M. Kashima, and H. Tourpe (2017). "Fintech and financial services: initial considerations." International Monetary Fund Staff Discussion Note SDN/17/05.

International Finance Corporation (2013). "Assessing Private Sector Contributions to Job Creation and Poverty Reduction." January, available at https://www.ifc.org/wps/wcm/connect/0fe6e2804e2c0a8f8d3bad7a9dd66321/IFC_FULL+JOB+STUDY+REPORT_JAN2013_FINAL.pdf?MOD=AJPERES (accessed December 17, 2018).

International Monetary Fund (2018). "The Bali Fintech Agenda." October, available at https://www.imf.org/en/Publications/Policy-Papers/Issues/2018/10/11/pp101118-bali-fintech-agenda (accessed December 17, 2018).

Jagtiani, J., and C. Lemieux (2017). "Fintech Lending: Financial Inclusion, Risk Pricing, and Alternative Information." July, available at https://www.fdic.gov/bank/analytical/cfr/bank-research-conference/annual-17th/papers/14-jagtiani.pdf (accessed January 2, 2019).

James, G., D. Witten, T. Hastie, and R. Tibshirani. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.

Jordan, M.I., and T.M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects, *Science*, *349*(6245), 255-60.

Kaffenberger, M., E. Totolo, and M. Soursourian (2018). "A Digital Credit Revolution: Insights from Borrowers in Kenya and Tanzania, CGAP Working Paper. October, available at https://www.cgap.org/sites/default/files/publications/Working-Paper-A-Digital-Credit-Revolution-Oct-2018.pdf (accessed January 2, 2019).

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The tyranny of data? the bright and dark sides of data-driven decision-making for social good. In *Transparent data mining for big and small data* (pp. 3-24). Springer, Cham.

Liberti, J.M., and M.A. Petersen (2018). "Information: Hard and Soft" (No. w25075), National Bureau of Economic Research.

Liu, T.Y. (2009). Learning to rank for information retrieval, *Foundations and Trends® in Information Retrieval*, *3*(3), 225-331.

Loterman, G., I. Brown, D. Martens, C. Mues, and B. Baesens (2012). "Benchmarking regression algorithms for loss given default modeling," *International Journal of Forecasting*, *28*(1), 161-70.

Lukonga, M.I. (2018). "Fintech, Inclusive Growth and Cyber Risks: Focus on the MENAP and CCA Regions," International Monetary Fund Working Paper No. 18/201.

McNeill, B.R., M. Saeed, E. Boukherouaa, and A. Tiffin (forthcoming). "Artificial Intelligence and Machine Learning: Promises and Challenges," International Monetary Fund Staff Discussion Note.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2018). *Foundations of machine learning*. MIT Press.

Mullainathan, S., and J. Spiess (2017). "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, *31*(2), 87-106.

Nazemi, A., K. Heidenreich, and F.J. Fabozzi (2018). "Improving corporate bond recovery rate prediction using multi-factor support vector regressions," *European Journal of Operational Research*, *271*(2), 664-75.

Philippon, T. (2016). "The fintech opportunity" (No. w22476), National Bureau of Economic Research.

Qi, M., and X. Zhao (2011). "Comparison of modeling methods for loss given default," *Journal of Banking & Finance*, *35*(11), 2842-55.

Qiu, J., Q. Wu, G. Ding, Y. Xu, and S. Feng (2016). "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 67.

Rajan, U., A. Seru, and V. Vig (2010). "Statistical default models and incentives," *American Economic Review*, *100*(2), 506-10.

Risk.net (2018). "OK, computer? Hurdles remain for machine learning in credit risk." July 3, available at https://www.risk.net/risk-management/5738771/ok-computer-hurdles-remain-for-machine-learning-in-credit-risk (accessed March 13, 2019).

Russell, S.J., and P. Norvig (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited.

Sahay, R., M. Čihák, P. N'Diaye, A. Barajas, R. Bi, D. Ayala, Y. Gao, A. Kyobe, L. Nguyen, C. Saborowski, K. Svirydzenka, and S.R. Yousefi (2015). "Rethinking Financial Deepening: Stability and Growth in Emerging Markets. International Monetary Fund Staff Discussion Note: SDN/15/08.

Sahay, R., M. Cihak, P. N'Diaye, A. Barajas, S. Mitra, A. Kyobe, Y.N. Mooi, and S.R. Yousefi (2015). "Financial Inclusion: Can It Meet Multiple Macroeconomic Goals?" International Monetary Fund Staff Discussion Note SDN/15/17.

Schweitzer, M.E., and B. Barkley (2017). "Is 'Fintech' Good for Small Business Borrowers? Impacts on Firm Growth and Customer Satisfaction." January, available at https://www.federalreserve.gov/conferences/files/is-fintech-good-for-small-business-borrowers.pdf (accessed March 14, 2019).

Scikit-learn (2018). *Scikit-learn user guide.* December, available at https://scikit-learn.org/dev/_downloads/scikit-learn-docs.pdf (accessed December 17, 2018).

Schmidhuber, J. (2015). "Deep learning in neural networks: An overview." *Neural networks*, *61*, 85-117.

Sirignano, J., A. Sadhwani, and K. Giesecke (2016). "Deep learning for mortgage risk," *arXiv preprint arXiv:1607.02470*.

Tirole, J. (2010). *The theory of corporate finance,* Princeton University Press.

Töws, E. (2016). *Advanced Methods for Loss Given Default Estimation* (Doctoral dissertation, Universität zu Köln).

Varian, H.R. (2014). "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, *28*(2), 3-28.

Wells Fargo website. "Know what lenders look for," Available at https://www.wellsfargo.com/financial-education/credit-management/five-c/ (accessed December 14, 2018).

World Economic Forum (2015). "The Future of FinTech: A Paradigm Shift in Small Business Finance," White Paper. October, available at http://www3.weforum.org/docs/IP/2015/FS/GAC15_The_Future_of_FinTech_Paradigm_Shift_Small_Business_Finance_report_2015.pdf (accessed December 17, 2018).

World Economic Forum (2016). "The complex regulatory landscape for FinTech: an uncertain future for small and medium-sized enterprise lending," White Paper. November, available at http://www3.weforum.org/docs/WEF_The_Complex_Regulatory_Landscape_for_FinTech_290816.pdf (accessed December 17, 2018).

Wooldridge, J.M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060-1089.