mathematics

MDPI

Article

Identifying Hidden Factors Associated with Household Emergency Fund Holdings: A Machine Learning Application

Wookjae Heo [1], Eunchan Kim [2,*], Eun Jin Kwak [3] and John E. Grable [4]

[1] Division of Consumer Science, White Lodging-J.W. Marriot Jr. School of Hospitability & Tourism Management, Purdue University, West Lafayette, IN 47907, USA; heo28@purdue.edu
[2] College of Business Administration, Seoul National University, Seoul 08826, Republic of Korea
[3] Department of Accounting and Finance, University of Wisconsin-Green Bay, Green Bay, WI 54311, USA; kwake@uwgb.edu
[4] Department of Financial Planning, Housing, and Consumer Economics, University of Georgia, Athens, GA 30602, USA; grable@uga.edu
* Correspondence: eunchan@snu.ac.kr

Abstract: This paper describes the results from a study designed to illustrate the use of machine learning analytical techniques from a household consumer perspective. The outcome of interest in this study is a household's degree of financial preparedness as indicated by the presence of an emergency fund. In this study, six machine learning algorithms were evaluated and then compared to predictions made using a conventional regression technique. The selected ML algorithms showed better prediction performance. Among the six ML algorithms, Gradient Boosting, *k*NN, and SVM were found to provide the most robust degree of prediction and classification. This paper contributes to the methodological literature in consumer studies as it relates to household financial behavior by showing that when prediction is the main purpose of a study, machine learning techniques provide detailed yet nuanced insights into behavior beyond traditional analytic methods.

Keywords: financial preparedness; emergency fund; machine learning; consumer studies

MSC: 68T07; 68T09

Citation: Heo, W.; Kim, E.; Kwak, E.J.; Grable, J.E. Identifying Hidden Factors Associated with Household Emergency Fund Holdings: A Machine Learning Application. *Mathematics* **2024**, *12*, 182. https://doi.org/10.3390/math12020182

Academic Editors: Jing Yao, Xiang Hu and Jingchao Li

Received: 24 November 2023
Revised: 29 December 2023
Accepted: 3 January 2024
Published: 5 January 2024

1. Introduction

As is the case with nearly all fields of study that fall under the area of the social sciences, much of the body of knowledge in the field of consumer studies is based on statistical results from conventional data methodological approaches, with regression procedures dominating the way researchers attempt to describe variable relationships and explain phenomena. Traditional regression techniques are designed to identify the marginal effects of specified and pre-selected factors based on theory and the existing literature. Conventional analytical techniques have been refined over the past half-century to increase explanatory power; however, even with advancements, conventional approaches remain limited in their explanatory power. Factors that might be possibly related to an outcome of interest, but have not been reported in the literature or thought to be theoretically relevant, are generally excluded from subsequent analyses. This means that the amount of explained variance across a wide number and variety of consumer studies outcomes is inevitably limited.

Big data analytical techniques, which tend to be atheoretical, have increasingly gained traction across the social sciences to acquire a deeper understanding of human attitudes and behaviors. Machine learning (ML)—a type of artificial intelligence application—is both a field of study and an umbrella term that describes algorithms that are built in such a way that hidden layers of information can be identified through a learning process based on training data and computational proofs. ML approaches are intended to supplement

the role of researchers by showing that variables that might have once been discarded in previous studies or not included at all in an empirical analysis can add insight into describing and explaining outcomes.

The purpose of this study is to illustrate the use of ML from a consumer studies perspective to improve data descriptions when compared to a conventional regression approach. The outcome of interest in this study is a household's degree of financial preparedness as indicated by the presence of an emergency fund (i.e., a measure based on household liquidity). As will be discussed later in this paper, numerous researchers have examined factors associated with holding an emergency fund, explaining the components of emergency savings, and predicting which households are most likely to meet liquidity ratio guidelines. A unique feature of much of the existing literature is that regardless of the research purpose, analysts tend to use similar variables when describing and predicting household emergency funds. These variables have come to represent the basis of many consumer-focused financial recommendations. A cursory review of this literature suggests, however, that other variables or relationships among variables is needed to gain a more comprehensive understanding of consumer financial preparedness to improve prediction rates.

When asked, financial service professionals, financial counselors, and financial educators tend to agree that managing household emergency funds involves the ongoing management of interacting variables. This is one reason why ecological systemic theory is prominently mentioned as a key explanatory model when emergency fund analyses are conducted at the household level [1,2]. As previously mentioned, much of the existing research has primarily sought to understand emergency funds within the confines of economic or financial theories using a delimited number of factors such as financial status or sociodemographic variables (e.g., [3,4]). While such studies have contributed positively to the literature by reinforcing existing theories and research findings, they may overlook the potential relevance of variables highly pertinent to how households manage emergency funds in practice. Methodologically, this signifies the need for an approach centered on pattern recognition and classification, as opposed to the identification of linear relationships upon which conventional studies have been based (e.g., [3–5]). Consequently, the combination of ecological systemic theory, pattern recognition, and classification underscores the necessity to consider complex system science models [6,7]. Furthermore, in the context of the social sciences and economics, where complex system science models are gaining acceptance, there is a need for research in personal finance utilizing ML techniques [6,8].

This study adds to the existing literature in several important ways. First, it employs ML in the context of a consumer studies topic. While some prior attempts within the field have been made (e.g., [9–15]), these efforts have been limited in their ability to compare various ML methods comprehensively. Another limitation is that some prior studies have relied on macro, rather than micro or household, data, which produce outcomes that are disconnected from a household's actual financial management activities. Consequently, this study is one of the few initial attempts to explain emergency fund management by integrating various ML techniques at the household level.

Second, previous studies have been limited to the assessment of a few central variables, including financial factors and sociodemographic factors, when studying emergency funds (e.g., [3,4]); this study is more expansive. Specifically, the analyses conducted in this study relied on a diverse set of variables that align with the research objectives. For instance, in addition to financial and sociodemographic factors, this study introduces a broad array of variables, including financial education, psychological factors, COVID-19-related factors, distance to financial service providers, and types of loans. This approach aligns well with the strengths of ML, which are designed to enhance predictive capabilities by combining numerous variables when classifying and describing relationships [16]. This study carries the potential to discover meaningful variables that have been previously unnoticed in existing research by supplementing ML predictions with additional variables potentially related to the management of emergency funds at the household level.

Third, as mentioned earlier, previous studies have typically assumed that variable relationships are linear, even when this assumption may not be practically relevant. Rather than rely on a linear assumption, this study is premised on pattern recognition and classification, distinct from models based on linear assumptions. Specifically, this study utilizes six ML algorithms as complex systems science models. While the six ML methods in this study have been widely used in empirical studies, their application in comparison to traditional linear assumption-based analytical methods is limited, particularly in relation to personal finance and consumer studies topics.

In summary, this paper contributes to the methodological literature in consumer studies by showing that when prediction is the main purpose of analysis (i.e., for use when making policy, creating education interventions, and advice giving), conventional analytical techniques may not always be the best solution. ML incorporating a larger set of variables that accounts for interactions between and among factors can offer a more robust and powerful way to increase predictive validity. In this regard, the research questions associated with this study are (a) What is the optimal ML algorithm to predict the presence of an emergency fund? (b) How do ML predictions perform when compared to a conventional logistic regression analysis? and (c) What are the most important factors associated with holding an emergency fund when viewed with an ML algorithm lens?

This study consists of sub-sections to arrive at the answer to these questions and deliver contributory points. Section 2 includes a background discussion about emergency funds and the methodological background of ML. Section 3 introduces the empirical model based on the background and methodological review. Section 4 describes the data and measurements utilized in the ML and logistic models. Section 5 illustrates the findings from each ML and the logistic model. Section 6 discusses the results. This paper concludes by describing this study's implications in Section 7.

2. Background

2.1. Household Emergency Funds

The ability of households to pay for unexpected emergencies and situations associated with unanticipated unemployment is a topic of interest to those who study and research consumer issues [17]. Household financial ratio analysis originates in consumer studies research that began in earnest in the last two decades of the 20th century. Johnson and Widdows [18] are generally given credit for being the first to adapt traditional business valuation ratios for use with households [19]. The liquidity ratio, also known as the emergency fund ratio, appears prominently in the early literature as a marker of household financial preparedness. Prather and Hanna [20] were among the first to publish standards and norms associated with the liquidity ratio, which is defined as the number of months a household can viably meet expenses in an emergency. The most commonly applied liquidity ratio formula is: Liquid Assets/(Minimum Monthly Fixed + Monthly Variable Expenses). The ratio indicates the number of months a household can weather an emergency. According to Lytton et al. [19], a household's goal should be to maintain an emergency fund equal to three months of living expenses (see also [21]). Based on this guideline, it has been estimated that less than one-third of U.S. households can adequately meet a financial emergency [22].

Gaining a unified understanding of the factors associated with holding an emergency fund that meets the liquidity ratio guideline can be complicated. Hanna et al. [23] noted that savings can be influenced directly by a household's stage in the lifecycle, which implies that the role of certain variables in describing savings patterns may differ across the lifecycle. Lifecycle theory suggests that households that expect higher income uncertainty should allocate more assets to precautionary saving [24]. Beyond anticipatory behavior, the literature also indicates that a number of personal and household characteristics are associated with an adequately funded emergency account. Bi and Montalto [22] reviewed the literature and they found age, education, income, race/ethnicity, spending behavior, risk tolerance, a willingness to borrow, holding negative economic expectations, motivation, diversification

of household income, the presence of other savings (e.g., retirement accounts), home equity, and available lines of credit provide needed information when attempting to describe who does or does not hold an emergency fund. In their study, Bi and Montalto concluded that the ability to save was more important than documenting a need to save when explaining emergency fund holdings. Others have identified factors such as financial confidence and financial knowledge as important when explaining emergency fund saving behavior.

2.2. An Introduction to Machine Learning

As the previous discussion highlights, the literature describing the characteristics associated with household emergency fund holdings has a long and robust history. Almost all previous studies that have been undertaken to describe the characteristics associated with holding emergency funds have been conducted using conventional linear-based modeling techniques. What has emerged from this literature is a common set of factors that are thought to be associated with the decision to build and maintain emergency fund assets (see [22]). An important caveat when evaluating the existing literature is the general lack of a description of the effect sizes of significant variable associations and very little discussion regarding the degree of model-explained variance. A careful examination of existing studies shows that while all the models described in the literature are statistically significant, the amount of explained variance rarely exceeds 40%. This means that other variables (or variable relationships) that have yet to be identified or used in models contribute significant explanatory power. What these variables are or how these variables interact is yet unknown.

Researchers are increasingly using ML techniques because it is now known that artificial intelligence algorithms can provide a deeper insight into the mechanisms underlying human attitudes and behaviors. ML algorithms can be used to identify what are sometimes referred to as hidden layers within data. Within these hidden layers are functions that may not be linearly related to the outcome of interest but are, nonetheless, important when viewed holistically in combination with other variables in a network [6]. A now ubiquitous example illustrates how hidden layers and networks perform in practice. In this example, assume a researcher wants to understand how people identify faces when viewed as an image. When the researcher shows study participants extracts of a subject's face (e.g., one eye, a tooth, nose), the researcher finds that these independent factors fail to reach statistical significance and thus do not provide enough information to describe a face accurately. In this example, the researcher wrongly concludes that people fail to use some visual cues when creating descriptions. What a person actually does is compile, through hidden layers of analyses, all relevant snippets of information to derive an identification. A single viewpoint cannot provide enough information to build a valid description, nor can eliminating some pieces of information improve validity. Similarly, researchers relying solely on conventional linear statistical techniques may inadvertently dismiss variables as irrelevant or unimportant when describing or predicting a social science outcome. Some researchers may dismiss potential explanatory variables altogether. Like limited pictorial extracts used when describing a face, traditional analytical techniques rarely provide more than a rough outline of an outcome or phenomenon.

This is where ML adds explanatory power beyond what can be obtained from most conventional data analysis methodologies. Kudyba and Kwaitinetz [25] and Thompson [26] described ML as improving classification by identifying patterns within large datasets. ML is generally used when a project aims to improve predictions. As with any statistical approach, the reliability of ML protocols depends on the data source and how variables are coded [27]. Numerous ML algorithms and models have been proposed and tested over the past two decades. Examples of early ML approaches include Naïve Bayes, Linear Discriminant Analysis, logistic regression, *k*-Nearest Neighbors, decision trees, Supportive Vector Machine, adaptive boosting, and Gradient Boosting methodologies. It is important to note that ML approaches do not always outperform conventional approaches. When an outcome is measured continuously, linear, polynomial, lasso, and ridge regressions sometimes provide a more robust level of prediction compared to more complex ML

techniques. According to Abiodun et al. [28], however, the sophistication of ML approaches has increased exponentially over the past decade, resulting in increasingly higher levels of reliability and robust prediction levels.

In this study, six ML algorithms are introduced and tested using the Orange package with Python [29] and then compared to predictions made using a conventional regression technique. The algorithms evaluated in this study included (a) *k*-Nearest Neighbor (*k*NN), (b) Gradient Boosting, (c) Naïve Bayes, (d) Support Vector Machine (SVM), (e) Stochastic Gradient Descent (SGD), and (f) Neural Networks (NN) (for more information about these techniques, see [28,30–32]). By comparing these six ML techniques, this study adds to the consumer studies methodology literature by illustrating how hidden connections can bring new and interesting variable associations that describe and predict consumer attitudes and behaviors to light.

2.3. Methodological Background: Machine Learning (ML) Algorithms and Their Applications in Financial and Consumer Research

As noted above, six ML algorithms were tested in this study. More than one algorithm was chosen because the literature shows that each offers unique advantages and disadvantages. A particular ML algorithm may perform well when the outcome is financial distress or bankruptcy but perform less well when applied to a credit scoring situation. The following discussion reviews the six ML algorithms tested in this study.

2.3.1. *k*-Nearest Neighbor (*k*NN)

As the name implies, *k*NN utilizes instance-based learning as a classification tool [33,34]. Instance-based learning means that the algorithm utilizes the vector space (i.e., space between objects) model, which makes *k*NN different from other classification algorithms. Because it relies on the vector space model, *k*NN can be utilized with cross-sectional data [35]. Various approaches can be used when assessing vector space [36]. When the outcome variable is categorical, Hamming distance can be utilized as shown in Equation (1):

$$Hamming\ distance = \sum_{i=1}^{I} Int(x_i \neq y_i) \quad (1)$$

where i indicates each observation; I is a set of observations i; x_i and y_i are the predictor and the outcome value with ith observation. When the outcome variable is a continuous variable, Euclidean distance, using the root of squared differences among observed samples, can be applied [37], or the Manhattan distance, using the absolute value of differences, can also be utilized as shown in Equations (2) and (3).

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{I}(x_i - y_i)^2} \quad (2)$$

$$Manhattan\ distance = \sum_{i=1}^{I}|x_i - y_i| \quad (3)$$

The combination of predictors and the outcome can be shown as (x_i,y_i) where i means the ith observation from the data (i = 1, 2, 3, … I). By using ascending order of distance, the observations can be allocated on a matrix as $d(x_1,\ y_1) \leq \cdots \leq d(x_i,\ y_i)$, where d is the distance from Equations (1), (2), or (3). When the outcome variable is categorical, the most frequent occurrence indicates the highest probability of belonging to the category shown in Equation (4). By using the probability, the expected category of the outcome is the maximum value from Function (4), as indicated in Equation (5):

$$\widehat{p_k} = \frac{\sum_{i=1}^{I}(y_{i=k})}{\widetilde{i}} \quad (4)$$

$$\hat{y} = argmax\widehat{p_k} \tag{5}$$

where a predictor is a categorical variable from 1 to K, k means the kth category; $\widehat{p_k}$ is the probability to be founded; and i is observed as the optimal observation (ith). In the case that the outcome variable is a continuous variable, a certain number of observations are selected ($n = i$) from $d(x_1, y_1) \leq \cdots \leq d(x_I, y_I)$. The selected observations are utilized to calculate the inverse distance weighted average, which produces the predicted value of an outcome from Equation (6):

$$\hat{y} = \frac{\sum_{i=1}^{I} \frac{1}{d(x_i,x)} y_i}{i} \tag{6}$$

As a classification algorithm, kNN is widely used for forecasting underweighted regression conditions. When kNN is combined with fuzzy vectoring, Östermark [38] suggested that kNN can be a useful tool for detecting data outliers, specifically when forecasting using finance and economic datasets. Because of the usability of kNN when making forecasts, this classification method has been adopted in various financial studies [39]. For instance, Meng et al. [33] adopted kNN to predict internet financial risk. They found an optimal number of categories for internet financial institutions. Phongmekin and Jarumaneeroj [40] utilized various algorithms (i.e., logistic regression, decision trees, Linear Discriminant Analysis, and kNN) to forecast stock exchange returns in Thailand. They found that kNN offers the best performance when predicting stock returns.

2.3.2. Gradient Boosting

Gradient Boosting was introduced by Breiman [41], which was then merged with a regression algorithm developed by Friedman [42]. Gradient boosting is an ensemble modeling technique that combines classification and regression methods [42,43]. As the term 'boosting' implies, weak patterns from a dataset can be strengthened through a learning process when the goal is to find the highest probability of prediction [38]. 'Gradient' means an error from each strengthened stage gradually decreases until the lowest error level is reached [44]. The basic learning process begins by measuring the error (i.e., residuals) between a predicted value and an observed value [45], as shown in Equation (7), which is called a loss function:

$$l(y_i, f(x_i)) = \frac{1}{2}(y_i - f(x_i))^2 \tag{7}$$

where i is the ith observation. The negative gradient format of Equation (7) produces residuals like those in Equation (8), which is a derivative of $l(y_i, f(x_i))$:

$$-\frac{\delta(y_i, f(x_i))}{\delta f(x_i)} = y_i - f(x_i) \tag{8}$$

As shown in Equation (8), the negative gradient produces a function similar to that of a regression residual (i.e., the difference between the predicted outcome and the actual outcome), which is how the name Gradient Boosting originated. Until the residuals are minimized, Gradient Boosting is iterated to make weak learners be combined, as shown in Equation (9):

$$\hat{y} = f(x) = \sum_{k=1}^{K} L_k + e \tag{9}$$

where k indicates each predictor; K is the optimal number to minimize the residual; and L_k is each different weak learner. Usually, the weak learner is a tree model developed using a predictor.

In practice, there are multiple types of Gradient Boosting, including categorical Gradient Boosting, scikit-learn Gradient Boosting, Extreme Gradient Boosting, and Extreme Gradient Boosting with random forest. Categorical Gradient Boosting utilizes features as categories [46]. Scikit-learn gradient boosting is a type of Gradient Boosting algorithm offered in Python (https://scikit-learn.org/stable/ accessed on 1 November 2023), whereas

Extreme Gradient Boosting is the most recent version of Gradient Boosting [9,47]. Each method was evaluated in this study.

The use of Gradient Boosting fits well with the research of interest in this study. Gradient Boosting is an ensemble model, which makes it particularly useful when conducting finance and business analyses [10,15]. Consider the work of Zhang and Haghni [15]. They utilized Gradient Boosting to improve travel time prediction in the transportation business. Specifically, they compared autoregressive integrated moving averages, random forest, and Gradient Boosting and concluded that Gradient Boosting showed better performance prediction. Guelman [10] investigated loss costs from Canadian insurers by comparing Gradient Boosting and a generalized linear model. Gradient Boosting was found to offer better performance in terms of prediction. Gradient Boosting has also been utilized in credit analyses. For instance, Chang et al. [44] compared various ML algorithms (i.e., group method of data handling, logistic, SVM, and Extreme Gradient Boosting). They observed Extreme Gradient Boosting to have outstanding performance when predicting credit risk. The approach has also been used to predict financial distress. Liu et al. [45] compared logistic, random forest, NN, SVM, and Gradient Boosting and noted that Gradient Boosting outperformed financial distress predictions. Carmona et al. [9] found the most impactful factors associated with bank failures using Gradient Boosting. Specifically, they compared bank failure prediction performance across logistic, random forest, and Extreme Gradient Boosting. They noted that Gradient Boosting provided the most meaningful insight when understanding bank failures.

2.3.3. Naïve Bayes

As the name implies, Naïve Bayes relies on Bayes' theorem; sometimes researchers refer to the approach as Bayes or independent Bayes [48]. In practical applications, Naïve Bayes is useful for clustering and classification purposes [49]. All variables or features in a prediction model are assumed to be independent [50]. Naïve Bayes utilizes conditional probability modeling by combining various predictors ($X_k \ni x_1, x_2, \ldots, x_k$) with a set of probabilities ($p(C_m|X_k)$), where k is the number of predictors and m means the number of probabilities found. Because Naïve Bayes assumes the independence of all predictors, the maximized probability of having a certain value (or category) can be found using Equations (10) and (11):

$$p(C_m|X_i) = \frac{1}{Z} p(C_m) \prod_{k=1}^{K} p(x_k|C_m) \tag{10}$$

$$\hat{y} = argmax_{m \in \{1,\ldots,M\}} p(C_m) \prod_{k=1}^{K} p(x_k|C_m) \tag{11}$$

Some researchers have criticized the approach because the independent assumption is unnatural and unrealistic [51]. This is the reason that the approach is termed naïve. However, because of the assumption of independence, Naïve Bayes offers a mathematical transformation advantage, making the dataset analysis more predictable [51].

Naïve Bayes has been utilized in various financial studies as a classification algorithm. Jadhav et al. [12] compared the efficacy of SVM, *k*NN, and Naïve Bayes as algorithms to predict credit ratings. After comparing the algorithms, they concluded that Naïve Bayes performed best. Deng [52] utilized Naïve Bayes to detect fraudulent financial statements in auditing. Deng noted that Naïve Bayes can provide unique insights. Similarly, Viaene et al. [14] utilized Naïve Bayes to detect financial fraud (i.e., consumers' faulty insurance claims). They concluded that the approach can improve prediction rates. Naïve Bayes has also been utilized in text classifications, such as when conducting a financial news analysis. Shihavuddin et al. [53] collected news articles about the Financial Times Stock Exchange 100 (FTSE100). Using Naïve Bayes, they concluded that not only does Naïve Bayes improve classification, but the approach can also be used to predict stock prices.

2.3.4. Support Vector Machine (SVM)

SVM classification is based on the concept of a hyperplane, which combines two separate classes [30]. The easiest way to understand classification by SVM is that a hyperplane is drawn among total samples. By drawing the hyperplane, two separate groups can be identified (e.g., upper and lower hyperplanes) as shown in Equations (12) and (13):

$$y = 1,\ when\ [B\sum x_k + a] > 0 \tag{12}$$

$$y = -1,\ when\ [B\sum x_k + a] < 0 \tag{13}$$

where k means each predictor and a is the constant in each hyperplane. Because of the complexities built into most datasets, the hyperplane is generally not well specified. Therefore, SVM sets the hyperplane by considering the maximum margin, the nearest vector from the potential hyperplane [54]. To draw a hyperplane when the maximum margin is found (Max M), SVM secures the optimal prediction performance. The function is shown in Equation (14), where B and a are assumed to be 1.00:

$$Max\ M,\ where\ y_k\left(B\sum x_k + a\right) \geq M \tag{14}$$

In addition to a hyperplane and maximum margin in SVM, kernel functioning is often used to help classify samples when the dataset and vectors are highly dimensional [54]. Because one straight hyperplane cannot easily be optimally identified when the dataset is highly dimensional, different types of hyperplanes can be utilized, including linear (i.e., straight), polynomial, radial basis function (RBF), and sigmoid. These types function in the hyperplane, called a kernel [30]. In the current study, four types of kernels were utilized.

SVM has been utilized widely in credit risk studies [55]. For example, the approach has been employed to predict credit scores [56,57]. Baesens et al. [58] compared various algorithms (i.e., SVM, logistic, discriminant analysis, kNN, Neural Networks (NN), and decision trees) to predict credit scores. They found that SVM and NN showed the best prediction performance compared to the other algorithms. Yang [59] introduced an adaptive credit-scoring system using a kernel-based SVM. Yang noted that the non-linear feature of datasets can be managed through kernel transformation. Kim and Ahn [60] utilized various ML algorithms (i.e., multiple discriminant analysis, multinomial logistic analysis, case-based reasoning, and an artificial neural network) to examine corporate credit rates. They found that SVM outperformed in detecting multiclass classification of corporate credit ratings. Similar findings have been reported by Chaudhuri and De [61], Chen and Hsiao [62], and Hsieh et al. [63] when making bankruptcy and financial distress predictions.

2.3.5. Stochastic Gradient Descent (SGD)

SGD emerged as an extension of previous theories, including the theory of adaptive pattern classifiers [64,65]. SGD is primarily used to help with data classifications. SGD begins by minimizing the errors (i.e., residuals) between predicted and observed values [66]. Specifically, SGD employs multiple iterations to minimize the errors in each gradient step [67] using Equation (15):

$$\theta = \theta - \eta\nabla_\theta J(\theta) \tag{15}$$

where θ is the parameters of all networks from predictors; $J(\theta)$ is the loss function by using θ; and η is the size of the learning rate. By repeating Equation (15), the parameters to minimize the value of the loss function can be estimated. SGD is popular because it is mathematically tractable and scalable [67]. Researchers like SGD because it helps solve optimization issues through stochastic approximation [68]. Because SGD relies on minimizing errors, regularization needs to be considered. Ridge and lasso are popular regularizations [69]. Elastic regularization can also be utilized [70]. The SGD approach can be employed when pre-selection or the transformation of explanatory variables is required and in situations where predictive machine learning scenarios are needed. The technique showcases robustness against outliers, as the steepest gradient algorithm emphasizes

the correct classification of data points closely aligned with their true labels. As such, SGD extends beyond a mere method for optimizing objective functions with appropriate smoothness properties. SGD applies to a diverse set of machine learning prediction methods (e.g., [71,72]).

Similar to the other ML algorithms, SGD has been used in various consumer and finance studies. Deepa et al. [69] utilized SGD to predict the early onset of diabetes. Compared to logistic models, SGD showed a better prediction outcome. Using SGD algorithms, they noted that SGD can be used to enhance prediction rates.

2.3.6. Neural Networks (NN)

NN is unquestionably the most mature of all algorithms within the ML area. NN offers flexibility when attempting to make classifications and when the goal of a project is to engage in future pattern recognition [25,26]. The uniqueness of NN is the approach's use of neurons as hidden layers. Neurons resemble the human brain architecture [73]. Because of the unique architecture, all inputs (i.e., features or variables) are assumed to be connected to all neurons. All neurons are also assumed to be connected to all expected outcomes [6]. The basic function of NN is shown in Equation (16):

$$y = a(\sum_{k=1}^{K} w_k x_k + e) \tag{16}$$

where k denotes the predictors; w_k is each predictor's weight; and a is e bias like the error terms. Because of the complex connectivity through neurons between inputs and outcomes, NN can be expected to improve the prediction rate of outcomes. For instance, if five variables are used as inputs to predict two outcomes, employing four neurons, then there are 20 connections between the five variables and four neurons and an additional eight connections between the four neurons and two outcomes. This interconnectedness means 160 possible pathways from the five variables to the two outcomes through the four neurons. As this example illustrates, neurons make all connectivity from inputs to outcomes so that the prediction of outcomes can be improved.

The first step when conducting an NN analysis is to define the optimal number of neurons. Because NN can employ any possible number of neurons, the number of neurons should be tested first to find the best performing model [74]. In this study, the number of neurons was first tested, and then the optimal number of neurons was employed in the final model.

As noted above, NN is a very popular ML technique. NN has been utilized to predict credit scores and other consumer behaviors. Baesens et al. [58] compared various algorithms, including SVM, logistic, discriminant analysis, *k*NN, NN, and decision trees, to conclude that SVM and NN show the best prediction performance compared to the other algorithms. Some researchers have utilized NN to detect financial fraud (e.g., fraud reporting, fraudulent use of credit cards, fraudulent financial statements, fraud claims) (e.g., [74–76]), whereas others have utilized NN for the prediction of bankruptcy and financial distress [57,77,78]. Heo et al. [11] applied NN to predict the savings-to-income and debt-to-asset ratios among U.S. households. They compared the prediction accuracy between NN and conventional regression models and found that NN provides a deeper and more meaningful insight into the savings-to-income ratio and the debt-to-asset ratio.

2.3.7. Comparison Analysis

As alluded to in the preceding discussion, it is common for researchers to check whether ML algorithms enhance predictions by comparing outcomes to the results generated from a conventional analytic tool. When the outcome variable is binary, a logistic regression model [79] is most often the comparison. A logistic regression model can be estimated from Equation (17):

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = a + \sum_{k=1}^{K} x_k \tag{17}$$

where *k* denotes the predictors. This approach was taken in this study. Specifically, the ML algorithms' predictions were compared to those predictions made using a maximum likelihood logistic regression.

3. Empirical Model Flow

3.1. Research Purpose and Analysis Structure

The overarching purpose of this study was to determine which modeling technique offers the best prediction rate when describing the presence of an emergency fund. As noted above, this study employed and compared various ML algorithms. A four-step analytical process was used, and the steps are described below.

Step 1: Find the best parameters across the various ML algorithms

Multiple sub-algorithms exist within nearly all ML algorithms (Naïve Baynes is an exception). For instance, in terms of *k*NN, the Euclidean method and the Manhattan method can be used to measure distance. For Gradient Boosting, four sub-algorithms are widely used: categorical, Extreme, Extreme with random forest, and scikit-learn. In the case of SVM, the kernel can be assumed to be linear, polynomial, RBF, or sigmoid. Three sub-algorithms exist for SGD (i.e., elastic, lasso, and ridge). At this step of the analytical process, each sub-algorithm was tested. For the conventional analysis (i.e., logistic regression), three types of feature selection were utilized (i.e., all variables, forward stepwise selection, and backward stepwise selection).

In addition to sub-algorithms, each ML algorithm can be affected by internal settings (i.e., parameter settings). Based on the parameter setting, the same algorithm may exhibit different degrees of performance robustness [80]. To account for this possibility, this study tested different parameters for each algorithm. For *k*NN, normally, the number of neighbors can affect classification performance. Therefore, different numbers of neighbors (i.e., from 1 to 100) were employed and compared to find the best tuning for the *k*NN algorithm. Regarding Gradient Boosting, the learning rate may affect the algorithm's performance. As such, various learning rate settings (i.e., 0.10, 0.15, 0.20, 0.25, and 0.30) were employed and compared to find the best application. For SVM, cost values are known to affect classification performance. To account for this, different cost values (i.e., 0.10, 1.00, 5.00, 10.00, 50.00, and 100.00) were employed and compared. It is also known that in terms of SGD, the learning rate may affect the algorithm's performance. To deal with this possibility, various learning rate settings (i.e., 0.001, 0.005, 0.010, 0.050, and 1.000) were employed and compared. For NN, the number of neurons can change the algorithm's performance. Therefore, different settings of neurons (i.e., 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, and 100) were utilized and compared to find the best performance outcome. As shown in Figure 1 (Part A and Part B and Line a), the first step in the analysis involved selecting the best performing sub-algorithms and the best tuning for each algorithm.

Step 2: Find the best ML prediction algorithm among the various ML algorithms

It is important to note that assuming that one specific ML algorithm will ever show a dominant performance across predictions and classifications is unrealistic. Rather, by the topical issue type and the predictive dataset's nature, diverse ML algorithms can be expected to show better/worse prediction and classification performance [27]. Given the binary feature of the dependent variable in this study, various classification algorithms were selected, as explained above. As shown in part A with line b in Figure 1, the second step in the analytical process involved finding the optimal ML algorithm from the selected six ML algorithms. The best prediction performance was selected as the most appropriate for use within the dataset.

Step 3: Check whether ML accuracies are higher than those offered by a conventional analysis

Even if a selected ML algorithm shows excellent performance across tested ML algorithms, the prediction function may actually offer a lower level of prediction when compared to a conventional analytical technique like logistic regression. Therefore, the

third step involves comparing the prediction performance of the selected ML algorithm and the conventional analysis (see parts A and B with line b, Figure 1).

Step 4: Determine which factors are associated with holding an emergency fund

Assuming the selected ML algorithm performs better than the conventional analysis, the influencing rank of input factors can be found by evaluating algorithm outcomes. The influencing rank can be viewed similarly to the significant variable list from a regression model, or the rank can differ. By checking the similarity or differences between the rank of influencing factors (ML algorithm) and the significant factors (logistic regression), it is possible to establish variable importance and possible linkages across variables that can then be examined at a later date. This step in the analytical process is crucial because some variables that emerge from an ML algorithm may not be significant in a traditional sense. Therefore, as shown in Figure 1 (line c for both parts A and B), the final step involves checking the variable list generated from the ML algorithms and the logistic analysis.

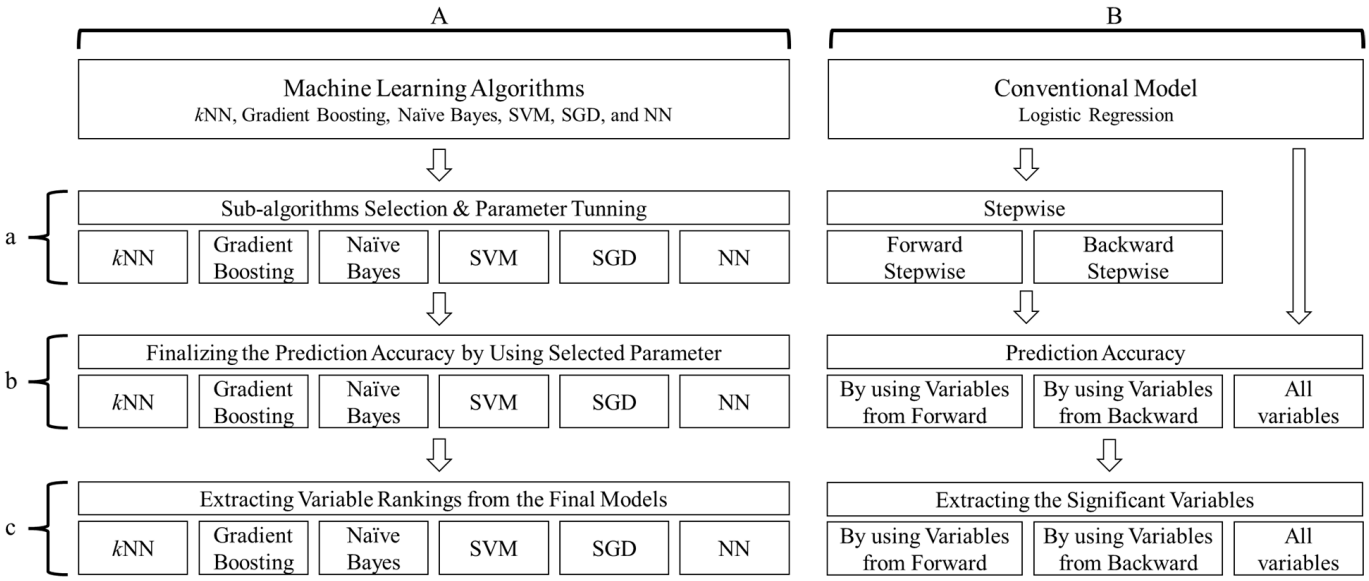

Figure 1. Analytic Structure for the Research (Abbreviations: *k*NN, *k*-Nearest Neighbor; NN, Neural Networks; SGD, Stochastic Gradient Descent; SVM, Support Vector Machine).

3.2. Analytic ML and the Conventional Analysis Process

Each ML algorithm test was conducted by dividing the sample into a training dataset and a test dataset. As shown in Figure 2, using the training dataset, each ML algorithm was used to identify the best prediction model. Data were split into training and testing datasets using a 50:50 random split ratio. As noted by Joseph [81], the split ratio varies by study and typically ranges from 80:20 division, 70:30, 60:40, and 50:50. The literature shows a conspicuous absence of definitive guidelines delineating the optimal or preferred data split ratio for a given dataset. As such, based on the comparatively small size of the dataset used in this study, the research team concluded that a 50:50 ratio was appropriate (see also [82,83]). Moreover, this ratio split allowed for robust validation of the data (i.e., k-fold validation). After a model was identified, the test dataset was utilized to validate the results from the test. If the model still showed a robust prediction outcome, the model was defined as optimal. The Python with Orange 3 visualization tool was used for all the tests. The conventional analysis utilized a similar procedure. A logistic regression model was estimated utilizing the training dataset. Results were validated using the test dataset. Stata 17.0 was used to estimate the models.

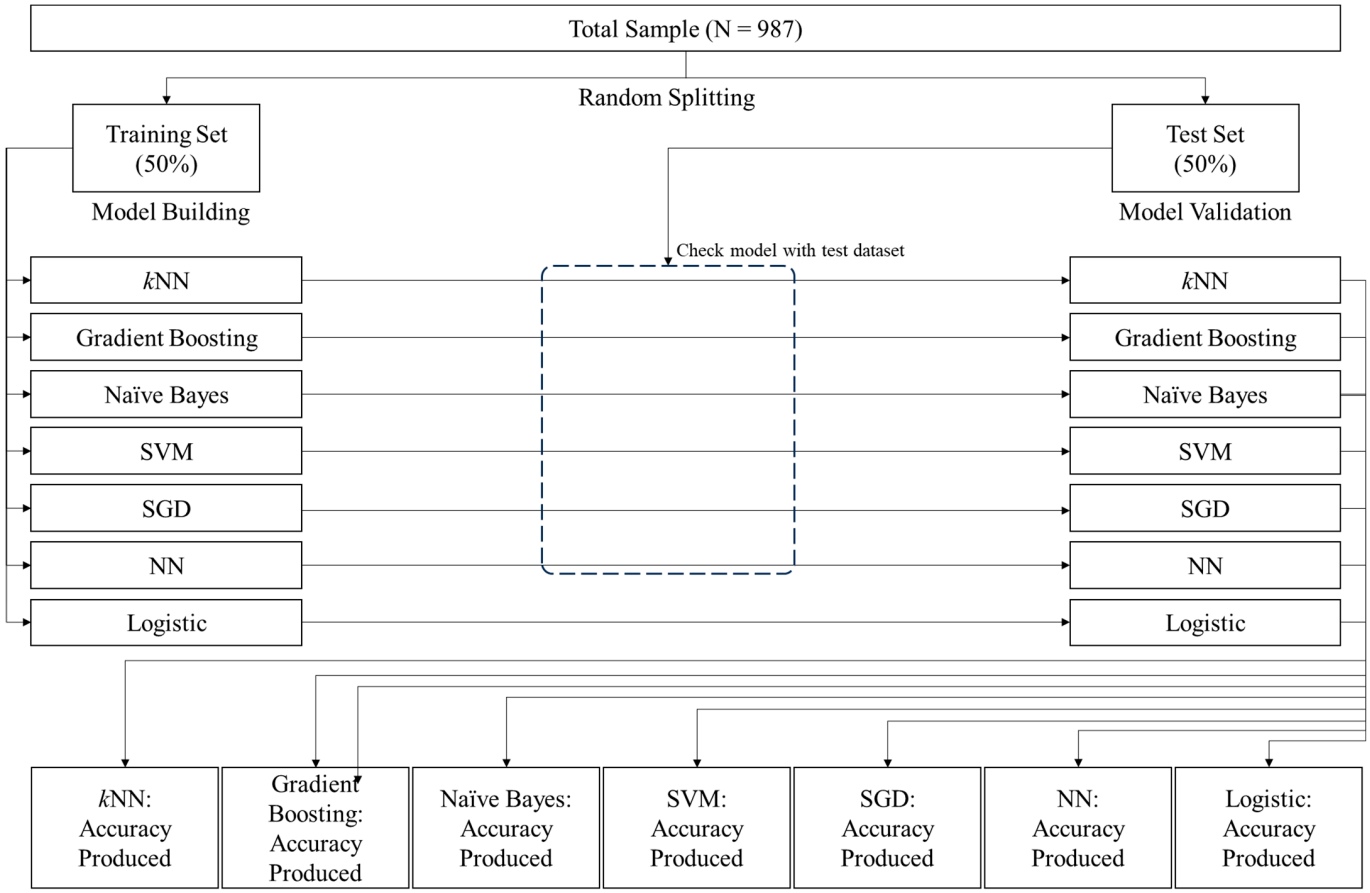

Figure 2. Analytic Process with ML Algorithms and Logistic Regression.

3.3. The Accuracy Estimation Method

To measure prediction accuracy, a receiver operator characteristics curve (ROC curve) and the area under the ROC curve (AUC) methodological approaches were utilized. An ROC curve is produced using two inputs: a true positive (TP) rate and a false positive (FP) rate [84]. The TP rate is calculated as the ratio between positive (i.e., correct) classifications and total positives. The FP rate is calculated using the ratio between negative (i.e., incorrect) classifications and total negatives. This indicates a more precise estimate when the TP rate is close to 1.00. The approach is also more precise when the FP rate is close to zero. An ROC curve shows the TP rate on the vertical axis and the FP rate on the horizontal axis. When an ROC curve shows a convex shape upward to the left, the accuracy is considered to be more precise. Additionally, the area under the curve is called the AUC, which indicates the power of the ROC (i.e., measured as 0.00 to 1.00) [44]. If the ROC curve has a vertical axis with a TP rate (i.e., zero to 1.00) and a horizontal axis with a FP rate (i.e., zero to 1.00), the area can be calculated from zero (zero times zero) to 1.00 (one times one).

3.4. The Factor Ranking Method

In Step 4, the rank of variables, in terms of prediction, is represented numerically (i.e., RReliefF). Whereas predictors in a logistic analysis can be evaluated using significance/insignificance estimates and marginal effects (i.e., coefficients), identifying high-ranking predictors using ML algorithms is more complex. For example, in the case of NN, all input variables connect to the outcome variable through neurons. Multiple weights are connected between a particular input variable and the outcome variable. There is not a specific number. As such, the evaluation of ML algorithms tends to focus on the complex combinations of input factors and the effects of variables on an outcome variable instead of the unique association between an input variable and the outcome variable.

For this study, variable ranks were identified using RReliefF. RReliefF is an advanced version of Relief [85] and ReliefF [86], which are generally accepted attribute estimators. Relief is the baseline of RReliefF. Robnik-Šikonja and Kononenko [87] introduced RReliefF, which was developed from Relief. The diff function, as shown below, can be used to better understand the baseline of RReliefF. The diff function is used to measure the distance among instances, which can be used to identify the nearest neighbors [87]. Equation (18) is used for categorical attributes, and Equation (19) is for continuous attributes:

$$diff(A,\ I_1,\ I_2) = \begin{cases} 0; value\ (A,\ I_1) = value\ (A, I_2) \\ 1;\ otherwise \end{cases} \quad (18)$$

$$diff(A,\ I_1,\ I_2) = \frac{|value\ (A,\ I_1) - value\ (A, I_2)|}{\max(A) - \min(A)} \quad (19)$$

These equations are used when investigating a dataset that comprises multiple examples, denoted as I_1, I_2, ..., I_n, situated within an instance space. Each example is characterized by a set of attributes, represented as A_i, where attributes are associated with each example. By using the diff function, the weight (W) of attribute A can be estimated as Relief by following Equation (20) [86]

$$W[A] = P(diff.value\ of\ A \mid nearest\ \ instance\ from\ diff.class) - P(diff.value\ of\ A \mid nearest\ \ instance\ from\ same\ class) \quad (20)$$

Based on the fundamental Relief framework, regressional ReliefF was introduced using Equation (21):

$$W[A] = \frac{P(diff.response \mid diff.value\ of\ A\ and\ nearest\ intances)P(diff.value\ of\ A \mid nearest\ instances)}{P(diff.response \mid nearest\ instances)} - \frac{(1-P(diff.response \mid diff.value\ of\ A\ and\ nearest\ intances))P(diff.value\ of\ A \mid nearest\ instances)}{1-P(diff.response \mid nearest\ instances)} \quad (21)$$

Compared to other attribute estimators (e.g., the root mean of squared error and mean absolute error), the RReliefF estimator uses a factor measured by considering interactions with other factors. RReliefF measures a factor's estimator contextually. A higher RReliefF number for a specific variable indicates that the factor is expected to predict the outcome with better (optimized) performance. Therefore, in this study, RReliefF was used to check the factors' ranking.

4. Data and Measurement

4.1. Data

Data were collected in 2021 using an online survey distributed in the United States. A survey agency invited 5900 consumer households to participate in this study; 1000 respondents answered all the questions; however, 13 respondents provided inaccurate information (e.g., reporting two years old for their age), which resulted in a useable sample of 987. Descriptive information for the sample is shown in Appendix A Table A1.

4.2. Measurement

The outcome variable was whether a respondent held an emergency fund or not. The variable was coded dichotomously (Have = 1; Not have = 0) based on an answer to the following question, "*Have you set aside emergency or rainy day funds that would cover your expenses for three months, in case of sickness, job loss, economic downturn, or other emergencies?*".

The input variables (i.e., predictors) were split into the following five categories in alignment with [88] and [89]: (a) financial statements and resources, (b) financial literacy and education, (c) psychological factors, (d) demographic factors, and (e) COVID-associated factors (used to account for the period of data collection).

The following binary-coded variables comprised the financial statements and resources category: (a) have auto loan or not; (b) have student loan or not; (c) have farm loan or not; (d) have equity loan or not; (e) have mortgage loan or not; (f) own house or not; (g) have saving account or not; (h) have checking account or not; (i) own term life insurance or not;

(j) own whole life insurance or not; (k) ever use payday loan; and (l) have health insurance or not. In addition, a categorical variable was included to account for the possibility of receiving financial advice for making financial decisions (i.e., 1 = have; 2 = do not know; 3 = no). Finally, a respondent's physical distance from their closest financial professional was asked and coded as follows: 1 = less than 5 miles; 2 = 5 to 10 miles; 3 = 10 to 20 miles; 4 = 20 to 50 miles; 5 = over 50 miles; and 6 = n/a or do not know.

Three variables comprised the financial literacy and education category: (a) had financial courses in high school (1 = Yes; 0 = otherwise); (b) had financial courses in college (1 = Yes; 0 = otherwise); and (c) objective financial literacy. The objective financial literacy variable was based on answers to three true/false questions [90], resulting in scores that could range from 0 (no correct answers) to 3 (all correct answers).

The psychological factors category was comprised of the following variables: (a) financial risk tolerance; (b) financial satisfaction; (c) financial stress; (d) financial self-efficacy; (e) locus of control; (f) life satisfaction; (g) the Rosenberg self-esteem scale; and (h) job insecurity. Financial risk tolerance was assessed using the Grable and Lytton's risk-taking propensity scale [91]. Scores ranged from 13 to 42. Financial satisfaction was measured using seven items on a five-point scale (min = 7; max = 35) (see [92]). Financial stress was measured using 24 items on a five-point scale (min = 24; max = 120) (see [88]). Financial self-efficacy was measured using six items, also on a five-point scale (min = 6; max = 30) (see [93]). Locus of control was measured using seven items on a five-point scale (min = 7; max = 35) (see [94]). Higher scores were representative of an external locus of control. Life satisfaction was measured using seven items on a seven-point scale (min = 5; max = 35) (see [95]). Self-esteem was measured with Rosenberg's 10-item scale that was assessed using a four-point scale (see [96]). Finally, job insecurity was measured using seven items on a five-point scale (min = 7; max = 35) (see [97]).

Demographic factors included (a) a variable representing the region of the country where a respondent lived, (b) work status, (c) agricultural working status, (d) education level, (e) marital status, (f) gender, (g) age, (h) whether a respondent lived in an urban area, (i) ethnicity, (j) income level, (k) number of children in a respondent's household, and (l) perceived health status. The region represented a respondent's state of residence. Work status was coded categorically as 1 = Full-Time; 2 = Part-Time; 3 = Self-Employed; 4 = Homemaker; 5 = Full-Time Student; and 6 = Not Working. Agriculture working status was coded as a categorical variable (1 = farm; 2 = ranch; 3 = agri-business; and 4 = not working in agriculture). Education level was coded categorically as 1 = high school or lower; 2 = some college; 3 = college; and 4 = postgraduate. Gender was coded as female or otherwise. Marital status was coded as a binary variable (i.e., single or otherwise). Age was measured in years. Living in an urban area was coded categorically as follows: 1 = urbanized area of 50,000 or more people; 2 = suburban area, near urbanized area with at least 2500 and less than 50,000 people; and 3 = rural area, all population, housing, and territory not included within any urban areas). Ethnicity was coded as a categorical variable, where 1 = White or Caucasian; 2 = Hispanic or Latino/a; 3 = Black or African American; 4 = Asian; 5 = Pacific Islander/Native American or Alaskan Native; and 6 = Other. Income level was coded categorically as 1 = Less than USD 15,000; 2 = USD 15,000 to USD 25,000; 3 = USD 25,000 to USD 35,000; 4 = USD 35,000 to USD 50,000; 5 = USD 50,000 to USD 75,000; 6 = USD 75,000 to USD 100,000; 7 = USD 100,000 to USD 150,000; and 8 = Over USD 150,000. The number of children living in a respondent's household was measured as a reported number. Finally, the perceived health status of a respondent was measured as a categorical variable (i.e., 1 = Excellent; 2 = Good; 3 = Fair; and 4 = Poor).

Finally, COVID factors were measured with items that asked how a respondent was affected by the COVID-19 virus and pandemic, how long a respondent expected the COVID-19 pandemic to last, and the receipt and timing of a stimulus check. The following items were used to evaluate perceptions of the COVID-19 pandemic: (a) how a respondent's financial situation was affected by COVID-19; (b) how a respondent's health condition

was affected by COVID-19; (c) how a respondent's general well-being was affected by COVID-19; and (d) how a respondent's work–life balance was affected by COVID-19. Answers were coded as 1 = almost no impact to 4 = serious impact. Perceptions about the duration of the pandemic were assessed by asking if (a) my financial situation will get better, get worse, or stay the same in three months; (b) my financial situation will get better, get worse, or stay the same in six months; or (c) my financial situation will get better, get worse, or stay same in one year. Answers were coded as 1 = get better; 2 = get worse; or 3 = stay the same. The timing of receiving a stimulus check was measured nominally as 1 = get stimulus check in April; 2 = get stimulus check in May; 3 = get stimulus check in June; 4 = get stimulus check in July; 5 = get stimulus check after July; 6 = do not know; 7 = do not want to answer; 8 = had not received stimulus check yet; and 9 = not eligible for a stimulus check.

5. Results

5.1. Identify the Best Parameters among the Various ML Algorithms

The first step in the ML analyses began by finding the best parameters and tuning the algorithms. Across the six ML algorithms, various parameters were tested and tuned. The tuning procedure is shown in Appendix B.

5.2. Results for Step 2: Find the Best ML Prediction Method among the Various ML Algorithms

It was determined that *k*NN and NN overfit the data somewhat. For example, the prediction accuracy (AUC) of both algorithms were strong when the models were built; however, the prediction accuracy was weakened when tested. Gradient Boosting offered the best performance with categorical consideration and a learning rate of 0.10 (see Table 1). However, *k*NN and SVM were still robust. Figure 3 shows the selected algorithms' ROC curves from the six ML algorithms.

Table 1. Prediction Accuracy Comparison across ML Algorithms.

ML	Selected Algorithm	Selected Parameter	Training	Test
*k*NN		Neighbor = 6	1.000	0.844
Gradient Boosting	Categorical	L.R. = 0.10	0.988	0.849
Naïve Bayes			0.871	0.818
SVM	Sigmoid	cost = 0.10	0.836	0.826
SGD	Lasso/Ridge	L.R. = 0.001	0.919	0.802
NN		Neuron = 30	1.000	0.793

Abbreviation: L.R., learning rate.

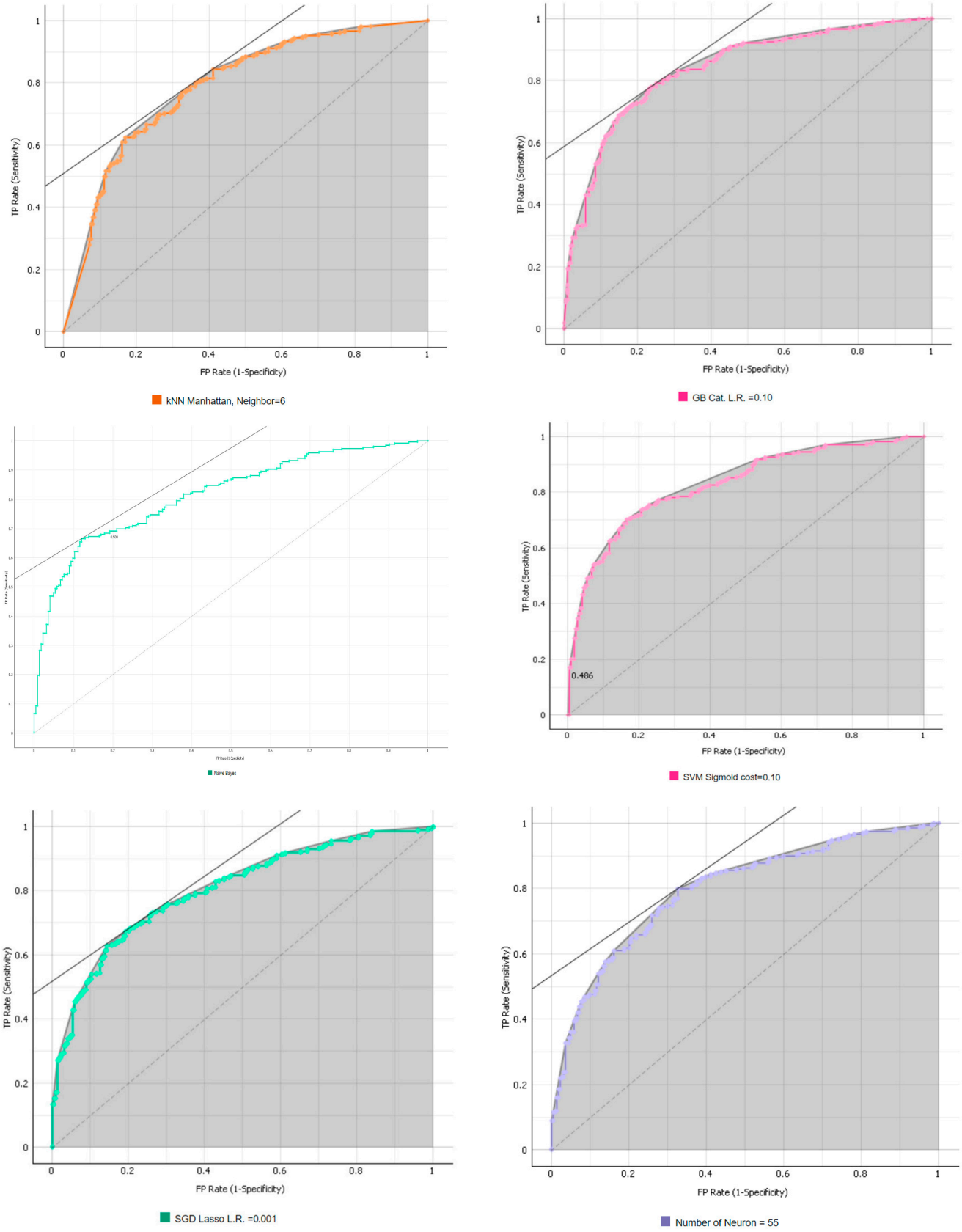

Figure 3. ROC Curves from the Best Predictions from Six ML Algorithms.

5.3. Results for Step 3: Check Whether the Accuracy of the ML Algorithms Is Higher Than the Accuracy Offered by a Logistic Regression

Table 2 shows the results from the logistic regression. As shown in Table 2, none of the variables had a significant effect in describing whether a respondent held an emergency fund. However, when the variables were added using a stepwise variable selection approach, several variables (i.e., savings account, mortgage loan, whole life insurance, no access to financial advisor, financial course in high school, financial satisfaction, financial self-efficacy, life satisfaction, number of children, and financial situation during the COVID-19 pandemic) were observed to be statistically significant.

Table 2. Logistic Regression Results (n = 475, 50% Random Splitting).

Variables	Logistic Regression with All Variables		Logistic Regression with Forward Stepwise		Logistic Regression with Backward Stepwise	
	Coefficient	*SE*	Coefficient	*SE*	Coefficient	*SE*
Auto loan	0.40	0.46				
Student loan	−0.61	0.47				
Farm loan	−0.04	0.80				
Equity loan	0.22	0.66				
Mortgage loan	−1.48	0.51			−0.69 *	0.31
Own house	0.50	0.51				
Saving acct.	−1.86	0.51	−1.40 ***	0.29	−1.28 ***	0.30
Checking acct.	−0.33	0.57				
Term L.I.	−0.08	0.41				
Whole L.I.	−1.02	0.51	−0.90 **	0.33	−0.81 *	0.34
FA do not know	−1.27	0.64				
FA no	−1.82	0.53	−1.12 ***	0.28	−1.14 ***	0.28
Payday loan	−0.66	0.56				
Health insurance	0.61	0.52				
FP Dist. 10 miles	0.49	0.56				
FP Dist. 20 miles	1.06	0.61				
FP Dist. 50 miles	1.08	0.91				
FP Dist. Over 50	1.40	1.19				
FP Dist. na	−0.21	0.57	−0.70 *	0.29	−0.80 **	0.30
Fin course in H.S.	−0.81	0.45	−1.01 **	0.29	−0.95 **	0.30
Fin course in Col.	−0.47	0.53				
Obj. Fin Knw.	−0.08	0.21				
Fin R.T.	0.04	0.05				
Fin Satisfaction	0.09	0.04	0.07 **	0.02	0.06 *	0.03
Fin Stress	0.02	0.01				
Fin Self-efficacy	−0.19	0.06			−0.08 *	0.03
L.O.C.	−0.05	0.05				
S.W.L.S.	0.08	0.03	0.08 ***	0.02	0.08 ***	0.02
Self-esteem	0.01	0.05				
Job insecurity	0.05	0.04				
WS Part-time	0.20	0.71				
WS Self-empl.	1.31	0.70				
WS Homemaker	−1.36	1.00				
WS Full stud.	0.28	0.82				
WS Not working	0.11	0.58				
Agri. Work	0.92	1.67				
Agri. R.Busi.	−0.77	1.02				
Agri. No.	−0.03	0.90				
Ed AA	0.44	0.50				
Ed BA	0.93	0.55				
Ed Grad.	0.66	0.74				
Single	0.22	0.45				
Female	0.12	0.41				
Age	0.02	0.02				

Table 2. *Cont.*

Variables	Logistic Regression with All Variables		Logistic Regression with Forward Stepwise		Logistic Regression with Backward Stepwise	
Suburban	0.42	0.44				
Rural	0.95	0.59				
Ethn. Hispanic	0.16	0.59				
Ethn. Black	−0.22	0.52				
Ethn. Asian	0.42	0.55				
Ethn. Pacific	−0.13	1.07				
Ethn. Others	−0.98	0.87				
Inc. 15 k to 25 k	−0.71	0.68				
Inc. 25 k to 35 k	−1.12	0.70				
Inc. 35 k to 50 k	−1.09	0.73				
Inc. 50 k to 75 k	−0.27	0.72				
Inc. 75 k to 100 k	−0.95	0.86				
Inc. 100 k to 150 k	−1.27	0.85				
Inc. > 150 k	1.26	1.27				
No. of Child	−0.66	0.20	−0.26 *	0.12	−0.27 *	0.12
Hth. Good	−0.24	0.49				
Hth. Fair	−1.17	0.68				
Hth. Poor	0.36	1.23				
Fin Situation	−0.48	0.23	−0.32 *	0.13		
H.Situation	−0.10	0.26				
WB.Situation	0.03	0.28				
Work. Situation	0.32	0.26				
3 months expect	−0.31	0.29				
6 months expect	0.14	0.27				
1 year expect	0.31	0.25				
Stim. May	0.58	0.77				
Stim. Jun.	−1.24	0.91				
Stim. Jul.	0.94	0.99				
Stim. Aft. Jul.	−0.74	0.66				
Stim. Dk	−0.72	0.72				
Stim. Na	−0.62	1.06				
Stim. No get	−1.10	0.74				
Stim. No elig.	−0.44	0.81				
Constant	8.47	3.90	3.93 ***	1.06	5.28 ***	1.25
R^2	0.54		0.41		0.41	
F	352.60		264.57 ***		268.99 ***	

Note. Reference group for auto loan, student loan, farm loan, equity loan, mortgage loan, own house, saving account, checking account, term life insurance, whole life insurance, financial course from high school, financial course from college are those who do not have them; male is the reference group for gender; ever had financial advice before is the reference group for experience of financial advice; distance to the accessible financial profession within 5 miles is the reference group for accessibility of financial professionals; full-time working status is the reference group for working status; working on a farm is the reference group for agriculture working status; high school or lower degree is the reference group for education level; living in urban area is the reference group for urban/suburban/rural living; lower than USD 15,000 is the reference group for income level; excellent health status is the reference group for health status; reference group for stimulus check is receiving stimulus check in April; the results for region (i.e., states) were omitted because the number of states and territories is too large to report while the sample size per location is too small. Significance level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Based on a sample size of 477, ROC graphs and AUCs (i.e., predictions made from the test dataset) are shown in Figure 4. The predictions resembled convex curves. The upper left ROC was made when all variables were included in the prediction; the lower left ROC was estimated when backward stepwise was utilized; the right upper ROC was made when forward stepwise was utilized.

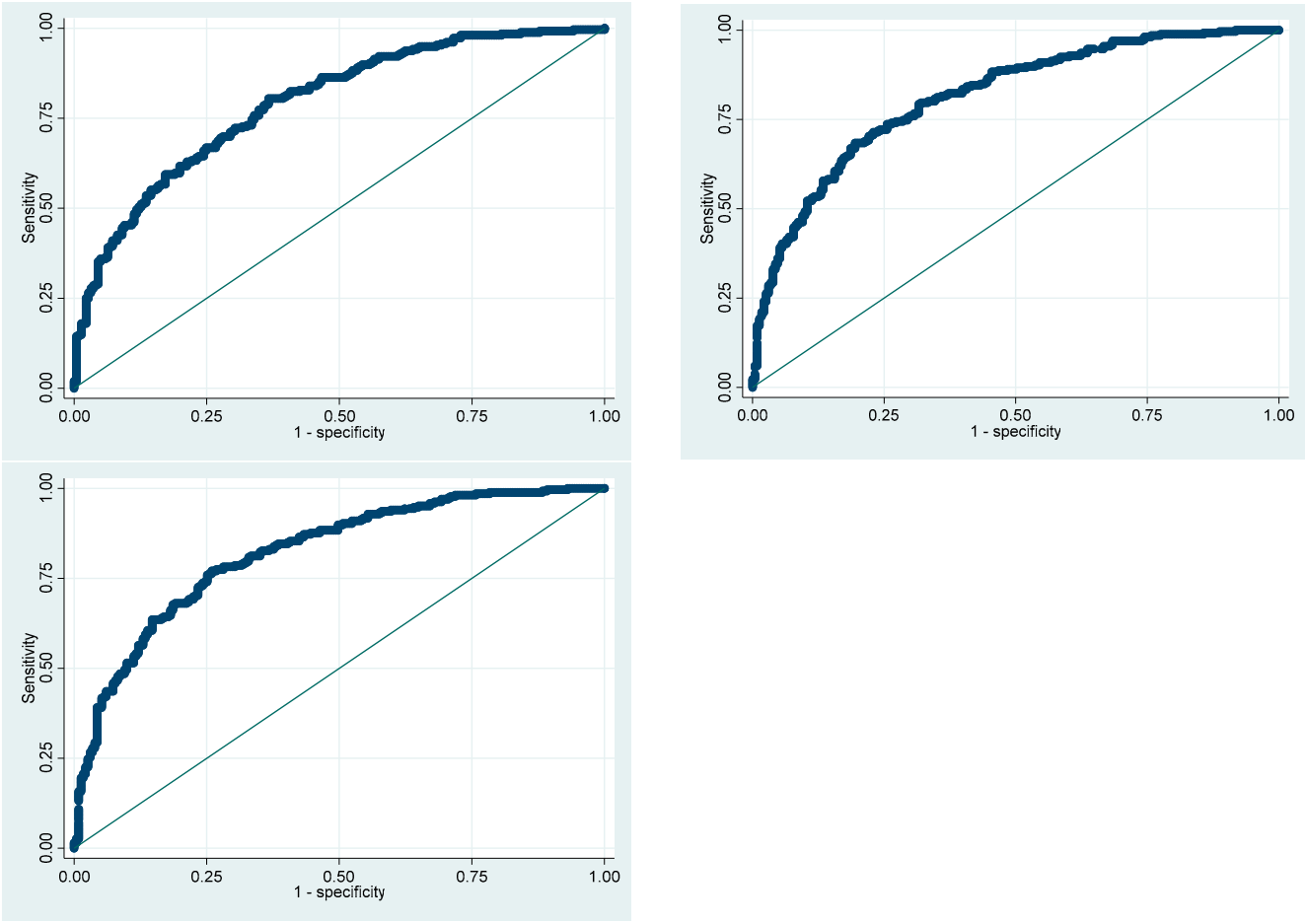

Figure 4. ROC Curves Based on Logistic Regression Modeling.

As shown in Table 3, AUC was under 0.800, which was lower than the ML AUC predictions. Even the worst performing ML exhibited a better AUC (i.e., 0.793 when ML was NN) compared to results from the logistic regression models (i.e., 0.754 when the variable list was determined via backward stepwise variable selection). This means that conventional analysis is proper when the research goal involves identifying significant variables; however, when the research goal involves maximizing prediction performance, ML algorithms provide a more robust insight into behavior (i.e., prediction accuracy can be maximized using ML techniques).

Table 3. AUC Comparison between ML Algorithms and Logistic Predictions.

ML	AUC from Test	Logistic Regression	AUC from Test
*k*NN	0.844	With all variables	0.703
Gradient Boosting	0.849	Forward stepwise	0.741
Naïve Bayes	0.818	Backward stepwise	0.754
SVM	0.826		
SGD	0.802		
NN	0.793		

Table 3 indicates that machine learning (ML) offers more (i.e., efficient) predictive performance than a logistic regression methodology. However, this does not necessarily mean that ML provides a better explanation. As previously explained, ML has the advantage of making better predictions by including more variables, as it incorporates the covariances inherent in each variable into a prediction. This means that some important features with

higher prediction weights are selected based on the covariance with other features. On the other hand, generalized linear models like logistic regression exclude covariances other than the unique covariance between an outcome and input variables. Traditional regression techniques focus on finding precise explanations for individual variables. This ultimately leads to an increase in explanatory power but a decrease in predictive power. Therefore, the results shown in Table 3 signify an improvement in the predictive power of ML but do not necessarily mean that the explanatory power of individual variables has improved.

For example, when looking at Table 2 (i.e., results from the logistic regression), variables that have a significant relationship with holding an emergency fund are easily identified. Most of these variables, including a household's financial situation, number of children, and holding a savings account, match with what has been reported in the previous literature. The explanatory power of these variables remains valid. However, Table 4 shows how different variables influenced these predictive performances. When comparing Tables 2 and 4, it becomes apparent that variables that were significant in Table 2 do not always have high predictive weights in Table 4. This indicates that in the case of the important variables shown in Table 4, various variables, as assumed by complex system science models and ecological system theory, contribute to better predictions. Therefore, the high predictive power in Table 3 and the variable rankings in Table 4 can play a role in identifying variables that conventional analyses, such as logistic regression, may overlook conceptually or theoretically. While ML may provide high predictive power, variables that were not statistically significant in the logistic regression (e.g., region, education level, financial self-efficacy, having a financial advisor, and farm loan) should be reconsidered as potentially important variables based on their high predictive weights, despite being overlooked in previous studies.

Table 4. Variable Rankings from Six ML Algorithms.

	***k*NN** Accuracy Rank = 2	**RF**	**GB** Accuracy Rank = 1	**RF**	**Naïve Bayes** Accuracy Rank = 4	**RF**	**SVM** Accuracy Rank = 3	**RF**	**SGD** Accuracy Rank = 5	**RF**	**NN** Accuracy Rank = 6	**RF**
1	Region	0.090	Education level	0.110	Fin Self-efficacy	0.075	Fin Course in Col.	0.176	Ever FA	0.128	Fin Course in Col.	0.136
2	Equity loan	0.080	Fin Course in Col.	0.104	Farm loan	0.070	Education level	0.158	Fin Course in Col.	0.108	Farm loan	0.134
3	Farm loan	0.076	Whole L.I.	0.102	Ever FA	0.069	Whole L.I.	0.158	Fin Course in H.S.	0.080	Ever FA	0.117
4	Fin Course in Col.	0.072	Region	0.089	Checking acct.	0.062	Farm loan	0.144	Single	0.078	Equity loan	0.102
5	Fin Course in H.S.	0.070	Ever FA	0.079	Fin Satisfaction	0.057	S.W.L.S.	0.115	Fin Satisfaction	0.074	Whole L.I.	0.088
6	Single	0.064	Farm loan	0.062	Region	0.054	Fin Satisfaction	0.112	Own house	0.072	Student loan	0.086
7	Ever FA.	0.061	Fin Satisfaction	0.061	Saving acct.	0.046	Ever FA	0.109	Gender	0.070	Payday loan	0.082
8	Education level	0.060	Gender	0.056	S.W.L.S.	0.044	Fin Stress	0.101	Farm loan	0.068	Education level	0.080
9	S.W.L.S.	0.054	Single	0.054	Payday loan	0.042	Fin Course in H.S.	0.092	Fin Self-efficacy	0.061	Fin Satisfaction	0.072
10	Payday loan	0.048	Fin Self-efficacy	0.053	Income level	0.040	Payday loan	0.088	S.W.L.S.	0.058	Term L.I.	0.064
11	Term L.I.	0.040	Income level	0.051	Age	0.035	Single	0.088	Fin Stress	0.057	S.W.L.S.	0.055
12	Fin Satisfaction	0.036	Mortgage loan	0.048	1 year expect	0.033	Agri. Work. Type	0.087	Dist. To. FP	0.046	Agri. Work. Type	0.051
13	Mortgage loan	0.034	Fin Stress	0.044	Fin Stress	0.028	Fin Self-efficacy	0.081	Obj. Fin Knw.	0.045	Auto loan	0.048
14	Health status	0.032	Own house	0.042	Education level	0.028	Term L.I.	0.076	Mortgage loan	0.044	Fin Self-efficacy	0.047
15	Fin Situation	0.031	Saving acct.	0.040	Stimulus	0.027	Checking acct.	0.070	Student loan	0.040	Fin Stress	0.046
16	Gender	0.028	Dist. To. FP	0.039	Fin Course in H.S.	0.026	Own house	0.070	Term L.I.	0.040	Saving acct.	0.044
17	Auto loan	0.028	Obj. Fin Knw.	0.035	WB.Situation	0.023	Fin Situation	0.067	Payday loan	0.034	Single	0.038
18	Income level	0.025	Equity loan	0.034	Equity loan	0.022	Health status	0.066	Agri. Work. Type	0.033	Ethnic	0.032
19	Fin Self-efficacy	0.023	6 months expect	0.032	Dist. To. FP	0.021	Work status	0.065	Region	0.033	WB.Situation	0.032
20	H.Situation	0.023	S.W.L.S.	0.031	Work status	0.019	Equity loan	0.064	Job insecurity	0.032	Fin Course in H.S.	0.032
21	Student loan	0.022	Job insecurity	0.031	Agri. Work. Type	0.019	H.Situation	0.059	Equity loan	0.032	Income	0.031
22	1 year expect	0.021	Term L.I.	0.030	L.O.C.	0.019	WB.Situation	0.059	Saving acct.	0.032	Checking acct.	0.030
23	Urban type	0.020	Agri. Work. Type	0.028	Auto loan	0.016	3 months expect	0.055	L.O.C.	0.031	Self-esteem	0.028
24	Agri. Work. Type	0.019	Fin Course in H.S.	0.024	6 months expect	0.016	L.O.C.	0.047	Education level	0.030	Work. Situation	0.026
25	Self-esteem	0.017	Ethnic	0.024	Health status	0.015	Obj. Fin Knw.	0.043	Age	0.030	Region	0.026
26	Fin Stress	0.014	Health status	0.022	Term L.I.	0.014	Work. Situation	0.041	WB.Situation	0.029	Fin Situation	0.023
27	Saving acct.	0.014	Payday loan	0.022	Fin R.T.	0.012	Stimulus	0.040	Income level	0.026	L.O.C.	0.023
28	Job insecurity	0.013	Auto loan	0.022	Obj. Fin Knw.	0.011	Income level	0.040	Self-esteem	0.025	Job insecurity	0.021
29	6 months expect	0.013	H.Situation	0.020	Single	0.010	Health insurance	0.040	Work status	0.024	Gender	0.020
30	Obj. Fin Knw.	0.012	L.O.C.	0.018	Urban	0.010	6 months expect	0.039	Health status	0.023	Mortgage loan	0.016
31	Work status	0.010	Age	0.016	H.Situation	0.010	Job insecurity	0.037	Urban type	0.021	Work status	0.013
32	Age	0.009	1 year expect	0.014	Self-esteem	0.007	Age	0.034	Health insurance	0.016	Age	0.010
33	Ethnic	0.008	Self-esteem	0.012	Fin Situation	0.007	Mortgage loan	0.034	1 year expect	0.015	Health status	0.009
34	Own house	0.006	No. of Child	0.005	Own house	0.006	Self-esteem	0.032	H.Situation	0.014	Fin R.T.	0.008
35	Health insurance	0.006	Fin R.T.	0.005	Job insecurity	0.003	Region	0.031	Fin Situation	0.011	H.Situation	0.008
36	L.O.C.	0.005	Checking acct.	0.004	Work. Situation	0.003	Student loan	0.030	Checking acct.	0.010	3 months expect	0.007
37	Stimulus	0.005	Fin Situation	0.003	Student loan	0.000	Saving acct.	0.028	Fin R.T.	0.009	Dist. To. FP	0.007
38	No. of Child	0.003	Work. Situation	0.000	No. of Child	−0.001	Auto loan	0.026	Auto loan	0.008	1 year expect	0.001
39	Whole L.I.	0.000	WB.Situation	−0.004	Ethnic	−0.009	Dist. To. FP	0.010	Work. Situation	0.008	No. of Child	0.000
40	Fin R.T.	−0.002	3 months expect	−0.005	3 months expect	−0.010	No. of Child	0.009	Stimulus	0.005	Own house	0.000
41	WB. Situation	−0.003	Health insurance	−0.006	Gender	−0.012	1 year expect	0.007	6 months expect	0.004	Obj. Fin Knw.	−0.002
42	Checking acct.	−0.012	Student loan	−0.010	Health insurance	−0.014	Gender	0.004	Whole L.I.	0.004	Urban type	−0.004
43	3 months expect	−0.025	Work status	−0.021	Mortgage loan	−0.018	Ethnic	−0.002	No. of Child	−0.004	Stimulus	−0.005
44	Work. Situation	−0.029	Urban type	−0.021	Whole L.I.	−0.024	Fin R.T.	−0.004	Ethnic	−0.013	Health insurance	−0.014
45	Dist. To FP.	−0.034	Stimulus	−0.030	Fin Course in Col.	−0.024	Urban type	−0.019	3 months expect	−0.020	6 months expect	−0.017

Abbreviations: Agri. Work. Type, agricultural working status; Dist. To. FP, distance to the financial professionals; Ever FA, ever have financial advice; GB, Gradient Boosting; RF, RReliefF; other abbreviations are same as shown in Table 2.

5.4. Results for Step 4: Determine Which Factors Are Associated with Holding an Emergency Fund

Table 4 shows the ranking importance of the best fitting ML algorithm (i.e., Gradient Boosting) across the variables evaluated in this study (i.e., RReliefF). Education level and having completed a financial course while in college ranked highly. This implies that educational attainment is important in helping someone gauge the need for an emergency fund. In addition, this indicates that promoting financial education, both in formal academic settings and through specialized courses, can be an effective strategy when encouraging individuals to (a) recognize the importance of emergency funds and (b) take proactive steps to establish emergency savings. Policy makers and educators should consider expanding financial education programs to enhance financial preparedness.

In addition, some financial-related psychological factors (i.e., financial satisfaction, financial self-efficacy, and financial stress) were found to be important. This implies that these factors are associated with holding an emergency fund. Financial institutions, financial service providers, and financial educators should incorporate psychological aspects into their financial literacy and counseling programs. Fostering financial satisfaction and self-efficacy while addressing financial stress is likely to help individuals develop positive emergency fund attitudes and behaviors.

Interestingly, COVID-19-related factors were not particularly important predictors in the model. This suggests that households are unlikely to change their emergency fund saving behavior even in the context of situational influences like a challenging economic situation.

Although Gradient Boosting was deemed to be the best model, the other ML algorithms produced comparable results. For instance, owning whole life insurance was an important variable when describing who holds an emergency fund across the model. This indicates that those who own whole life insurance are more concerned about their future self and the financial welfare of other household members (i.e., individuals who own life insurance generally exhibit a heightened awareness of their long-term financial security and the financial well-being of their family). Financial service providers can use this insight to emphasize the importance of comprehensive financial planning, including both insurance and emergency fund considerations. Similarly, educational factors (i.e., education level, completing a financial course in high school, or a financial course while in college) were found to be important predictors across the ML algorithms.

The ML results differed in significant ways from the logistic regression estimates. Compared to the Gradient Boosting model, taking a financial course in college and financial stress were unimportant in the logistic regression. Even so, there were some similarities. For instance, owning whole life insurance, taking a financial course in high school, and financial satisfaction ranked highly across the models. This indicates theoretical connections between these variables and holding an emergency fund. This study illustrates that combining insights from different analytical approaches can lead to a more comprehensive understanding and effective promotion of emergency fund savings.

6. Discussion

ML and big data analytical techniques have, over the past decade, garnered increasing attention among researchers, educators, and policy makers as a way to obtain deeper insight into social science phenomena. This study adds to the growing consumer studies methodological literature by illustrating how ML techniques can be applied to assessing household consumer attitudes and behaviors and how ML methods can improve prediction rates.

The outcome variable in this study was whether a household held an emergency fund, which was used to indicate a household's degree of financial preparedness. The existing financial ratio literature is relatively consistent in reporting that those who hold emergency savings share a common demographic profile [3,4]. They tend to have high income, are more educated, and have greater wealth. It is important to note, however, that nearly all profiles reported in the literature were constructed using traditional methodologies,

primarily regression techniques. At the outset of this paper, it was hypothesized that while existing profiles may remain valid, other variables might also be influential in describing who does and does not hold emergency savings. Traditional regression modeling techniques do not account for hidden layers between and among variables. While it is possible to create moderation and mediation models, to do so with large data is nearly impossible when the constraints associated with regression modeling are applied. This study's methodological approach dealt with this issue by showing that when prediction or profiling is the main purpose of a study, ML algorithms can provide a more nuanced insight into consumer behavior compared to more commonly used statistical analysis techniques [7,16].

This study compared and tested several ML algorithms to determine which offers the most robust prediction rate. The ML algorithm outputs were compared to estimates derived from logistic regression models. Several takeaways emerged from these analyses. First, those using ML techniques must know that parameter tuning is not optional. Incorrect parameter tuning results in lowered prediction and classification rates. Those who adopt ML algorithms in consumer studies should consider this point and compare tuning performance when conceptualizing studies. Second, sub-algorithms should be considered. Using an incorrect sub-algorithm will almost always lower prediction and classification validity. Third, when evaluating ML algorithm outputs, it is important to remember that ML algorithms do not show marginal effects. Instead, ML algorithms provide a ranked ordering of predictors. As such, the interpretation of an ML analysis should not be considered deterministic. Instead, the interpretation of an ML output needs to be conceptualized as more in line with an explorative introduction.

In this study, Gradient Boosting, *k*NN, and SVM were found to provide the most robust degrees of prediction and classification. Gradient Boosting offered the best prediction rate, which aligns with what others have reported in the literature (e.g., [9,10,15,44]). Gradient boosting is an ensemble modeling technique that integrates classification and regression methods [42,43]. The ensemble of classification and regression estimation works well when optimizing prediction accuracy [31] and minimizing error levels [44]. What is particularly interesting in this study is that income and wealth—factors generally considered the most descriptive of financial preparedness—were not highly ranked in the Gradient Boosting algorithm, nor with *k*NN or SVM. This insight differs from what is generally shown using regression techniques [3]. However, educational factors and the existence of financial obligations were more important. It appears that a consumer must possess the financial literacy to anticipate the need for emergency savings, formulate a plan to build an emergency fund, and implement the plan. The consumer must also have an objective reason to hold emergency fund assets. The existence of loans is one reason a consumer may opt to hold assets in an emergency fund. Likewise, a consumer needs to hold an attitudinal disposition that values one's future self or the well-being of household members. The consistently high ranking of life insurance in the ML algorithms suggests that the ability to plan for the future is an important characteristic among those holding emergency fund assets. The region variable in the *k*NN model is worthy of future research. The variable represents the state where a respondent resided at the time of the survey. It appears that some consumers are more likely than others to take financial preparedness steps. Specifically, those living in rural areas who also hold existing debt, are predicted to be more likely to hold an emergency fund.

This study represents a noteworthy advancement in consumer studies literature, particularly in the domains of personal finance and financial planning. This paper illustrates the value of ML techniques when predicting behavior. While numerous researchers have utilized ML methodologies with social science datasets (e.g., [9–15]), these efforts have sometimes suffered from limitations, such as their inability to comprehensively compare diverse ML methods or their focus on non-household factors. This means that the practical relevance of findings about household financial management has notable limitations. This

paper is one of the few studies to comprehensively analyze the nuances associated with holding an emergency fund at the household level.

Another significant contribution of this paper is the expanded scope of variables that were used to predict holding an emergency fund. Rather than rely on a limited set of preexisting variables as described in the literature (i.e., primarily financial factors and sociodemographic attributes) (e.g., [3,4]), this study introduced a broader range of variables, including financial education, psychological aspects, COVID-19-related factors, distance to financial service providers, and holding various types of loans. This approach aligns well with ML's capacity to leverage multiple variables [16], potentially unveiling overlooked variables that could significantly contribute to understanding the dynamics of emergency fund management.

Moreover, this study departs from the prevailing practice of assuming linear relationships between and among variables. The ML technique uses a pattern recognition and classification approach, making it possible to move beyond traditional linear assumptions. To achieve this, six distinct ML algorithms were employed as complex systems science models. The application of these algorithms allowed for a comprehensive investigation of the potential contributions of ML to the field of consumer studies. Notably, each ML algorithm underwent meticulous parameter tuning and calibration, extending beyond algorithmic utilization to demonstrate the application of ML techniques to address complex questions. The comprehensive approach in this study underscores the commitment to advancing the understanding of emergency fund management dynamics and enhancing the practical applicability of ML in consumer studies.

In summary, the results from this study advance the methodological body of literature for those working in the consumer studies field. This study shows that ML algorithms can be used to improve predictions and classifications of consumer attitudes and behaviors. Future research should align the results from this study with existing models and profiles of those who hold emergency savings. Information from such studies can be used by financial educators, consumer advocates, and policy makers when helping households achieve greater levels of financial preparedness.

7. Conclusions

This study is noteworthy in making significant theoretical, practical, and methodological contributions to consumer studies. The theoretical contribution lies in its application of ML techniques to the study of household financial decision making. Unlike traditional linear models, this study used a pattern recognition and classification methodology, shedding light on the intricate complexities underlying emergency fund management. The findings from this study challenge conventional beliefs by highlighting the importance of financial literacy, financial obligations, and a positive attitude towards future financial well-being as key factors in predicting who is more likely to hold emergency savings, with income and wealth taking a secondary role.

On a practical level, findings from this study underscore the critical importance of parameter tuning and sub-algorithm selection when employing ML techniques in consumer studies. This paper offers valuable insights into the use of ML algorithms when predicting and classifying consumer attitudes and behaviors, which can have direct applications for financial service providers, financial educators, consumer advocates, and policy makers. Moreover, this study expands the spectrum of variables considered, incorporating financial education, psychological factors, COVID-19-related variables, and others, thereby enhancing the predictive capacity of models to understand the dynamics of emergency fund management.

Even in the context of these significant contributions, limitations need to be acknowledged. ML techniques, while improving prediction rates, do not readily provide straightforward marginal effects. Thus, some researchers use ML algorithms as a starting point in identifying key variables for use in secondary models. While this study evaluated six robust ML algorithms, including Gradient Boosting, *k*NN, and SVM, further research is needed

to determine when one particular approach should be used to address a specific research question. Further advanced ML algorithms, such as Generative Adversarial Network, Recurrent Neural Network, or Convolutional Neural Network, should be evaluated in future studies. In the context of this study, additional research is needed to decipher regional variations in holding an emergency fund. Future studies should also aim to integrate the findings with existing models and profiles of emergency savings holders. Doing so will contribute to a better understanding of the financial preparedness of households. In addition, the current ML algorithms are all well-known algorithms. Even in the context of these limitations and opportunities for future work, this study advances the consumer studies methodological landscape by showcasing how ML techniques can enrich the field's comprehension of consumer attitudes and behaviors, particularly within the context of holding an emergency fund.

Author Contributions: Conceptualization, W.H. and E.K.; methodology, W.H. and E.K.; software, W.H. and E.K.; validation, W.H., E.K., E.J.K. and J.E.G.; formal analysis, W.H.; investigation, E.K.; data curation, E.J.K.; writing—original draft preparation, W.H., E.K., E.J.K. and J.E.G.; writing—review and editing, W.H., E.K., E.J.K. and J.E.G.; supervision, W.H. and J.E.G.; funding acquisition, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the USDA National Institute of Food and Agriculture, Hatch project 1017028 and Multistates project 1019968.

Data Availability Statement: The research dataset can be obtained upon a proper request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Descriptive Table (N = 987).

Category	Variable	Frequency	Percentage	Mean	SD
Outcome	Em. Fund (=Have)	538	54.51%		
	Auto loan (=Have)	355	35.97%		
	Student loan (=Have)	307	31.10%		
	Farm loan (=Have)	156	15.81%		
	Equity loan (=Have)	181	18.34%		
	Mortgage loan (=Have)	320	32.42%		
	Own house	487	49.34%		
Financial Factors	Saving acct.	650	65.86%		
	Checking acct.	807	81.76%		
	Term L.I.	418	42.35%		
	Whole L.I.	289	29.28%		
	FA have	330	33.43%		
	FA do not know	143	14.19%		
	FA no	514	52.08%		
	Payday loan	274	27.76%		
	Health insurance	776	78.62%		
	FP Dist. 5 miles	216	21.88%		
	FP Dist. 10 miles	229	22.29%		
	FP Dist. 20 miles	140	14.18%		
	FP Dist. 50 miles	67	6.79%		
	FP Dist. Over 50	44	4.46%		
	FP Dist. na	300	30.40%		
Financial Education	Fin course in H.S. (=Have)	363	36.78%		
	Fin course in Col. (=Have)	296	29.99%		

Table A1. *Cont.*

Category	Variable	Frequency	Percentage	Mean	SD
	Obj. Fin Knw.			1.56	1.00
	Fin R.T.			22.70	4.71
	Fin Satisfaction			22.54	7.31
	Fin Stress			66.95	27.71
Psych.	Fin Self-efficacy			15.59	5.22
Factors	L.O.C.			18.57	6.27
	S.W.L.S.			21.56	8.73
	Self-esteem			28.38	5.05
	Job insecurity			19.69	4.55
	WS Full-time	396	40.12%		
	WS Part-time	93	9.42%		
	WS Self-empl.	80	8.11%		
	WS Homemaker	59	5.98%		
	WS Full stud.	78	7.90%		
	WS Not working	281	28.47%		
	Agri. Farm	113	11.45%		
	Agri. Ranch	21	2.13%		
	Agri. R.Busi	66	6.69%		
	Agri. No	787	79.74%		
	Ed High	279	28.27%		
	Ed AA	269	27.25%		
	Ed BA	269	27.25%		
	Ed Grad.	170	17.22%		
	Single	503	50.96%		
	Female	501	50.76%		
	Age			38.86	15.29
	Urban	419	42.45%		
	Suburban	396	40.12%		
Demo.	Rural	172	17.43%		
Factors	Ethn. White	357	36.17%		
	Ethn. Hispanic	135	13.68%		
	Ethn. Black	250	25.33%		
	Ethn. Asian	149	15.10%		
	Ethn. Pacific	38	3.85%		
	Ethn. Others	58	5.88%		
	Inc. < 15 k	175	17.73%		
	Inc. 15 k to 25 k	118	11.96%		
	Inc. 25 k to 35 k	138	13.98%		
	Inc. 35 k to 50 k	127	12.87%		
	Inc. 50 k to 75 k	148	14.99%		
	Inc. 75 k to 100 k	98	9.93%		
	Inc. 100 k to 150 k	110	11.14%		
	Inc. > 150 k	73	7.40%		
	No. of Child			0.74	1.08
	Hth. Excellent	280	28.37%		
	Hth. Good	468	47.42%		
	Hth. Fair	190	19.25%		
	Hth. Poor	49	4.96%		
	Region	-	-		

Table A1. *Cont.*

Category	Variable	Frequency	Percentage	Mean	SD
	Fin Situation			2.33	1.08
	H.Situation			2.00	1.05
	WB.Situation			2.29	1.07
	Work. Situation			2.27	1.09
	3 months expect			2.06	0.90
	6 months expect			1.91	0.89
	1 year expect			1.72	0.88
C-19 Factors	Stim. Apr.	164	16.62%		
	Stim. May.	101	10.23%		
	Stim. Jun.	78	7.90%		
	Stim. Jul.	61	6.18%		
	Stim. Aft. Jul.	159	16.11%		
	Stim. Dk	133	13.48%		
	Stim. Na	39	3.95%		
	Stim. No get	129	13.07%		
	Stim. Not elig.	123	12.46%		

Abbreviation: Em. Fund, emergency fund; acct., account; L.I., life insurance; FA have, ever have financial advice; FA do not know, not knowing whether have financial advice; FA no, never have financial advice; FP Dist. 5 miles, financial professionals are accessible within 5 miles; FP Dist. 10 miles, financial professionals are accessible within 10 miles; FP Dist. 20 miles, financial professionals are accessible within 20 miles; FP Dist. 50 miles, financial professionals are accessible within 50 miles; FP Dist. Over 50, financial professionals are accessible over 50 miles; FP Dist. na, the accessibility of financial professionals is not known; Fin course in H.S., financial course from high school; Fin course in Col., financial course from college; Obj. Fin Knw., objective financial knowledge; Psych. Factors, psychological factors; Fin R.T., financial risk tolerance; Fin Satisfaction, financial satisfaction; Fin Stress, financial stress; Fin Self-efficacy, financial self-efficacy; L.O.C., locus of control; S.W.L.S., satisfaction with life scale; Demo., demographic; WS Full-time, working status as full-time worker; WS Part-time, working status as part-time worker; WS Self-empl., working status as self-employed; WS Homemaker, working status as homemaker; WS Full stud., working status as full- time student; WS Not working, working status as not working; Agri. Farm, working in agriculture as farm worker; Agri. Ranch, working in agriculture as ranch worker; Agri. R.Busi., working in agriculture as rural business; Agri. No., not working in agriculture; Ed High, education level as high school or lower; Ed AA, some college with associate degree; Ed BA, college with Bachelors' degree; Ed Grad., education level as graduate or higher degree; Ethn. White, ethnic group as White or Caucasian; Ethn. Hispanic, ethnic group as Hispanic or Latino(a); Ethn. Black, ethnic group as black or African American; Ethn. Asian, ethnic group as Asian; Ethn. Pacific, ethnic group as Pacific Islander, Native American, or Alaskan Native; Ethn. Others, ethnic group as others; Inc. < 15 k, income level as lower than USD 15,000; Inc. 15 k to 25 k, income level between USD 15,000 and USD 25,000; Inc. 25 k to 35 k, income level between USD 25,000 and USD 35,000; Inc. 35 k to 50 k, income level between USD 35,000 and USD 50,000; Inc. 50 k to 75 k, income level between USD 50,000 and USD 75,000; Inc. 75 k to 100 k, income level between USD 75,000 and USD 100,000; Inc. 100 k to 150 k, income level between USD 100,000 and USD 150,000; Inc. > 150 k, income level over USD 150,000; # Child, number of children in a household; Hth Excellent, health status as excellent health status; Hth Good, health status as good health status; Hth Fair, health status as fair health status; Hth Poor, health status as poor health status; C-19 Factors, COVID-19 factors; Fin Situation, the financial situation affected by COVID-19; H.Situation, the health situation affected by COVID-19; WB.Situation, general well-being affected by COVID-19; Work. Situation, work-balance affected by COVID-19; 3 months expect, the expected financial situation in 3 months; 6 months expect, the expected financial situation in 6 months; 1 year expect, the expected financial situation in 1 year; Stim. Apr., getting stimulus check in April; Stim. May., getting stimulus check in May; Stim. Jun., getting stimulus check in June; Stim. Jul., getting stimulus check in July; Stim. Aft. Jul., getting stimulus check after July; Stim. Dk, do not know whether get stimulus check or not; Stim. Na, do not want to answer; Stim. No get, the respondent did not get stimulus check; Stim. Not elig., the respondent is not eligible to get stimulus check.

Appendix B

ML Tuning: Identify the Best Parameters among the Various ML Algorithms

Tables A2–A7 and Figures A1–A6 show each ML algorithm's accuracy given the constraints of each algorithm's settings. In the case of *k*NN, both the Euclidean and Manhattan models showed robust predictions in the training dataset. However, when the models were checked using the test dataset, the Manhattan distance algorithm exhibited a better prediction rate. Regarding parameter tuning, the Manhattan model showed the best performance when there were three to eight neighbors. It was determined that the best parameter distance was six (6).

Table A2. Algorithm and Parameter Selection—*k*NN.

	Training		Test	
Number of Neighbors	**Euclidean AUC**	**Manhattan AUC**	**Euclidean AUC**	**Manhattan AUC**
1	1.000	1.000	0.686	0.835
2	1.000	1.000	0.742	0.834
3	1.000	1.000	0.754	0.840
4	1.000	1.000	0.775	0.840
5	1.000	1.000	0.779	0.840
6	1.000	1.000	0.785	0.844
7	1.000	1.000	0.786	0.838
8	1.000	1.000	0.786	0.842
9	1.000	1.000	0.786	0.838
10	1.000	1.000	0.794	0.836
20	1.000	1.000	0.809	0.828
30	1.000	1.000	0.810	0.825
40	1.000	1.000	0.807	0.818
50	1.000	1.000	0.811	0.809
60	1.000	1.000	0.802	0.806
70	1.000	1.000	0.803	0.799
80	1.000	1.000	0.801	0.776
90	1.000	1.000	0.799	0.708
100	1.000	1.000	0.795	0.834

Note. AUC represents the prediction accuracy of the model. AUC ranges in value from 0.00 to 1.00, and the higher the AUC, the better the model predicts. Abbreviation: AUC, area under the curve.

Figure A1 shows the representative ROC curves for *k*NN. The upper left graph is the ROC graph for the Euclidean model with 30 neighbors; the left lower graph is the ROC graph for the Euclidean model with 50 neighbors; the right upper graph is the ROC graph for Manhattan model with six neighbors; the lower right graph is the ROC graph for Manhattan model with eight neighbors. The dark section under the curve is the area used to calculate AUC. As shown in Figure A1, the ROC curves were convex, indicating that *k*NN performed well in prediction. The AUC was maximized when *k*NN was performed using the Manhattan model with six neighbors.

In the case of Gradient Boosting, the four sub-algorithms exhibited prediction robustness with the training dataset. However, when the algorithms were checked using the test dataset, categorical Gradient Boosting showed better prediction. Regarding parameter tuning, categorical Gradient Boosting showed the best performance when the learning rate was 0.10, as shown in Table A3.

Table A3. Algorithm and Parameter Selection—Gradient Boosting.

	Training				Test			
L.R.	**Cat. AUC**	**Ext. AUC**	**Ext. RF AUC**	**Scikit AUC**	**Cat. AUC**	**Ext. AUC**	**Ext. RF AUC**	**Scikit AUC**
0.10	0.988	1.000	1.000	0.968	0.849	0.842	0.842	0.836
0.15	1.000	1.000	1.000	0.981	0.835	0.840	0.840	0.838
0.20	0.998	1.000	1.000	0.985	0.827	0.840	0.840	0.842
0.25	1.000	1.000	1.000	0.991	0.838	0.834	0.834	0.833
0.30	0.999	1.000	1.000	0.994	0.833	0.838	0.838	0.829

Abbreviation: Cat., Categorical Gradient Boosting; Ext., Extreme Gradient Boosting; Ext. RF, Extreme Gradient Boosting with random forest; L.R., learning rate; Scikit, Scikit version of Gradient Boosting.

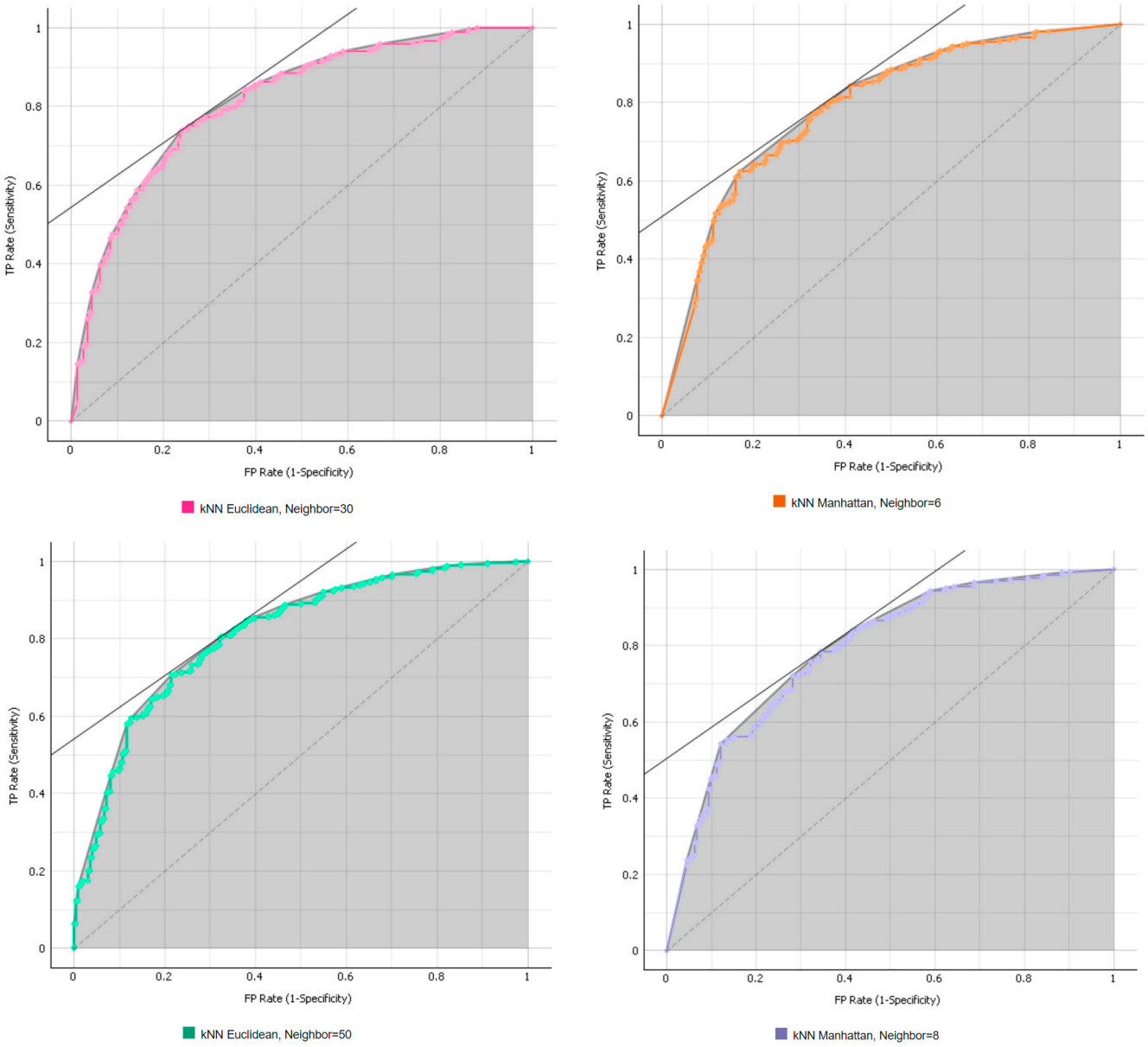

Figure A1. ROC Curves for Algorithm and Parameter Selection—*k*NN.

Figure A2 shows the representative ROC curves for the Gradient Boosting algorithms. The upper left graph is the ROC illustration for Categorical Gradient Boosting with a learning rate of 0.10; the left lower graph is the ROC graph for Extreme Gradient Boosting with a learning rate of 0.10; the right upper graph is the ROC graph for Extreme Gradient Boosting with random forest with a learning rate of 0.10; the lower right graph is the ROC graph for Scikit Gradient Boosting with a learning rate of 0.10. AUC was calculated using the dark area under the curve. As shown in Figure A2, the ROC curves were convex, suggesting that prediction was robust with Gradient Boosting. The AUC was the largest when Categorical Gradient Boosting was performed with a learning rate of 0.10.

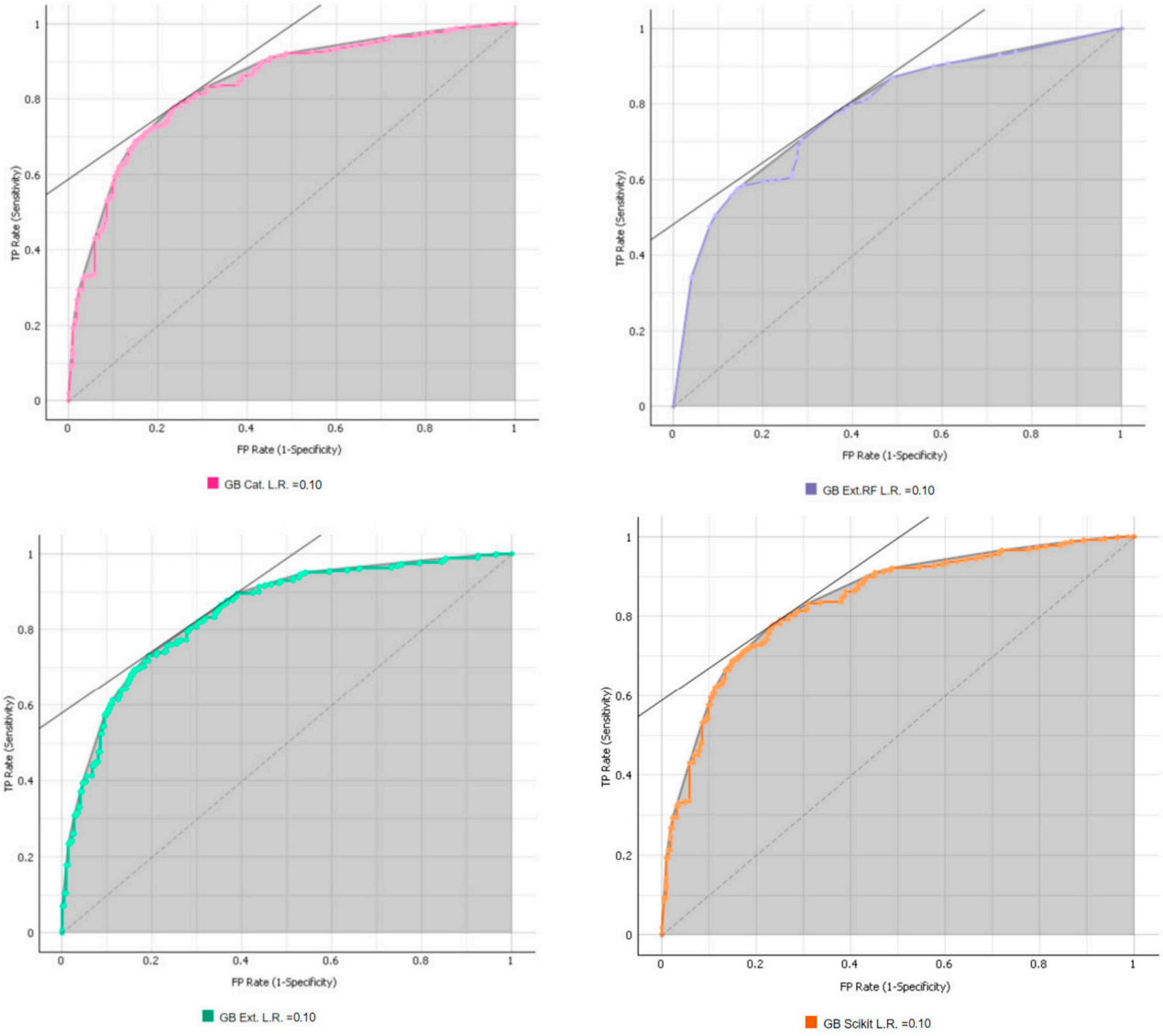

Figure A2. ROC Curve for Algorithm and Parameter Selection—Gradient Boosting.

There are no comparable sub-algorithms and parameter tuning estimates in the case of Naïve Bayes. Table A4 and Figure A3 show the Naïve Bayes' AUC and ROC curves. The dark area under the curve is the area used to estimate AUC.

Table A4. Algorithm and Parameter Selection—Naïve Bayes.

Training AUC	Test AUC
0.871	0.818

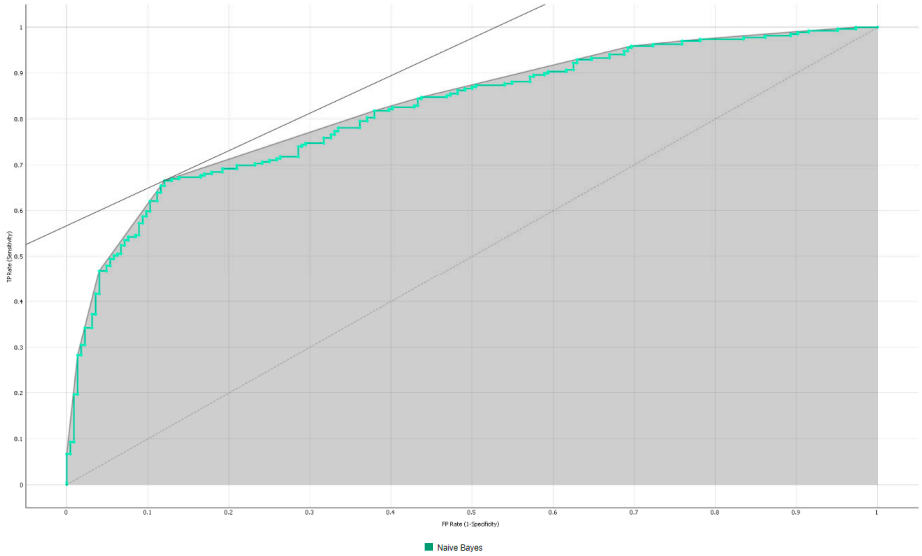

Figure A3. ROC Curve for Algorithm and Parameter Selection—Naïve Bayes.

Table A5 shows the Support Vector Machine (SVM) algorithm accuracy. In the case of SVM, the Radial Basis Function (RBF) kernel model exhibited the best prediction with the training dataset. As an optimal parameter setting, the cost was set between 5 and 100. However, when the algorithm was checked using the test dataset, optimal performance by RBF was overfit (i.e., better performance in training but worse performance when tested). It was determined that the sigmoid model was better in terms of prediction (i.e., the outcomes were similar between the training (0.836) and the test (0.826) datasets). The sigmoid kernel model with cost = 0.10 showed stable prediction (i.e., no overfitting issue) and optimal performance.

Table A5. Algorithm and Parameter Selection—SVM.

	Training				Test			
c	**Linear AUC**	**Poly. AUC**	**RBF AUC**	**Sigmoid AUC**	**Linear AUC**	**Poly. AUC**	**RBF AUC**	**Sigmoid AUC**
0.10	0.584	0.944	0.901	0.836	0.442	0.822	0.812	0.826
1.00	0.754	0.982	0.969	0.774	0.719	0.778	0.825	0.773
5.00	0.754	0.977	0.997	0.769	0.720	0.762	0.784	0.747
10.00	0.754	0.977	0.996	0.765	0.720	0.762	0.803	0.738
50.00	0.754	0.977	0.996	0.759	0.720	0.762	0.803	0.733
100.00	0.754	0.977	0.996	0.754	0.280	0.762	0.803	0.729

Abbreviation: c, cost; Linear, SVM with linear kernel; Poly., SVM with polynomial kernel; RBF, SVM with radial based function kernel; Sigmoid, SVM with sigmoid kernel.

Figure A4 shows the representative ROC curves for SVM. The upper left graph is the ROC graph for the linear SVM with a cost of 0.10; the left lower graph is the ROC graph for the polynomial SVM with a cost of 0.10; the right upper graph is the ROC graph for the RBF SVM with a cost of 0.10; the lower right graph is the ROC graph for the sigmoid SVM with a cost of 0.10. The dark area under the curve was used to calculate AUC. As shown in Figure A4, the ROC curves were convex, indicating that three of the SVMs performed well in prediction. When SVM was performed using a linear assumption, the prediction was suboptimal, as indicated by the concave graph. The AUC was optimized when SVM was performed with sigmoid with a cost of 0.10.

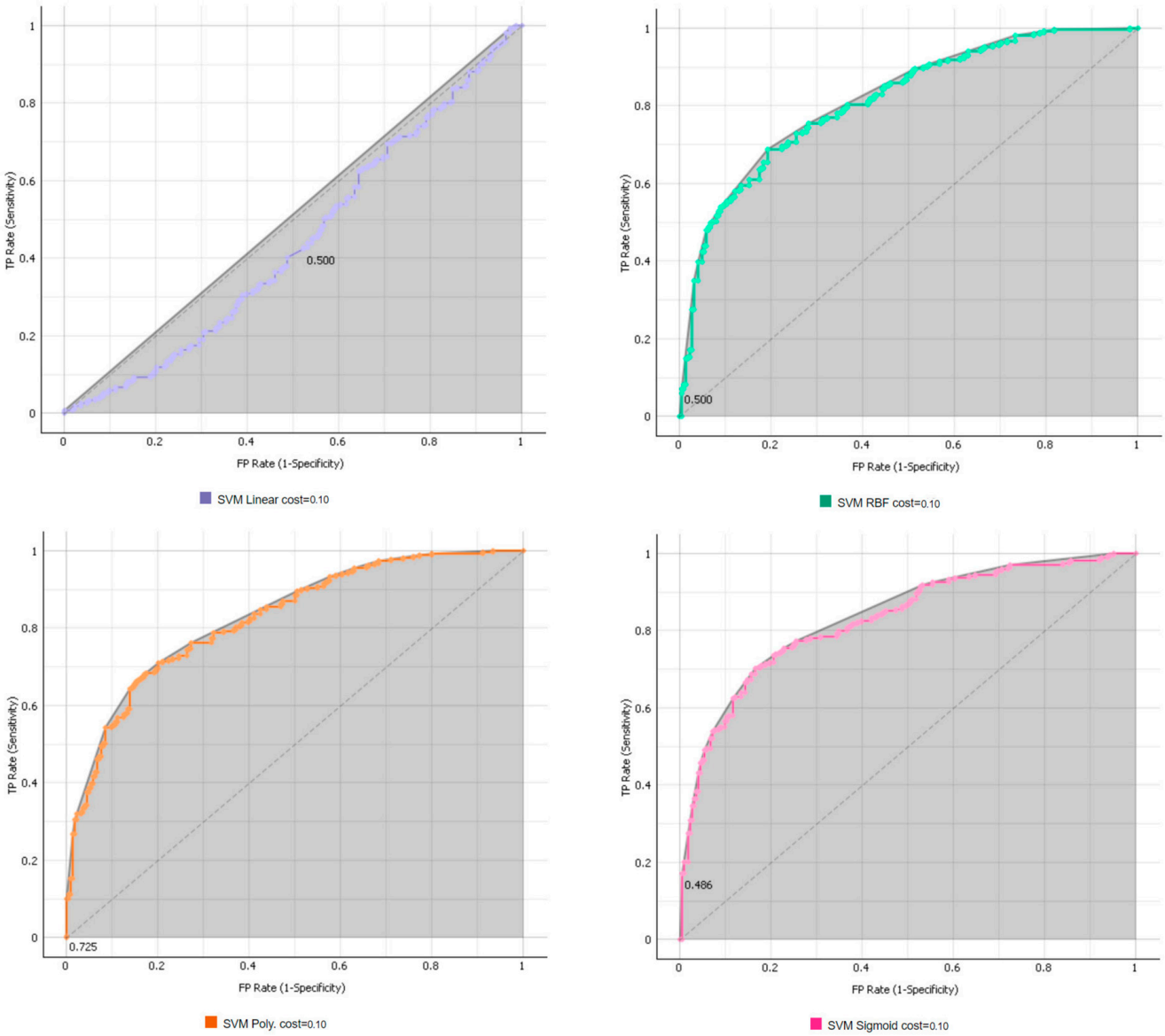

Figure A4. ROC Curve for Algorithm and Parameter Selection—SVM.

In the Stochastic Gradient Descent (SGD) shown in Table A6, reasonably good prediction rates were observed under three assumptions in the training dataset with learning rates of 0.001 and 0.005. However, when the algorithms were checked, the learning rate of 0.001 showed the best level of prediction. The type of assumption used when modeling did not lead to significant differences between the models as long as the learning rate remained at 0.001.

Table A6. Algorithm and Parameter Selection—SGD.

	Training			Test		
L.R.	**Elastic AUC**	**Lasso AUC**	**Ridge AUC**	**Elastic AUC**	**Lasso AUC**	**Ridge AUC**
0.001	0.919	0.919	0.919	0.801	0.802	0.802
0.005	0.924	0.924	0.924	0.790	0.786	0.785
0.010	0.923	0.922	0.922	0.778	0.780	0.787
0.050	0.896	0.896	0.895	0.713	0.759	0.770
0.100	0.870	0.890	0.877	0.759	0.774	0.659

Abbreviation: L.R., learning rate.

Figure A5 shows the representative ROC curves for SGD. The upper left graph is the ROC graph for lasso SGD with a learning rate of 0.001; the left lower graph is the ROC graph for ridge SGD with a learning rate of 0.001; the right upper graph is the ROC graph for lasso SGD with a learning rate of 0.05; the lower right graph is the ROC graph for elastic SGD with a learning rate of 0.001. As with the other analyses, the dark area under the curve was used to calculate AUC. As shown in Figure A5, the ROC curves were convex, indicating that each SGD performed well in prediction. The AUC was the largest when SGD was performed, with lasso/ridge with a learning rate of 0.001.

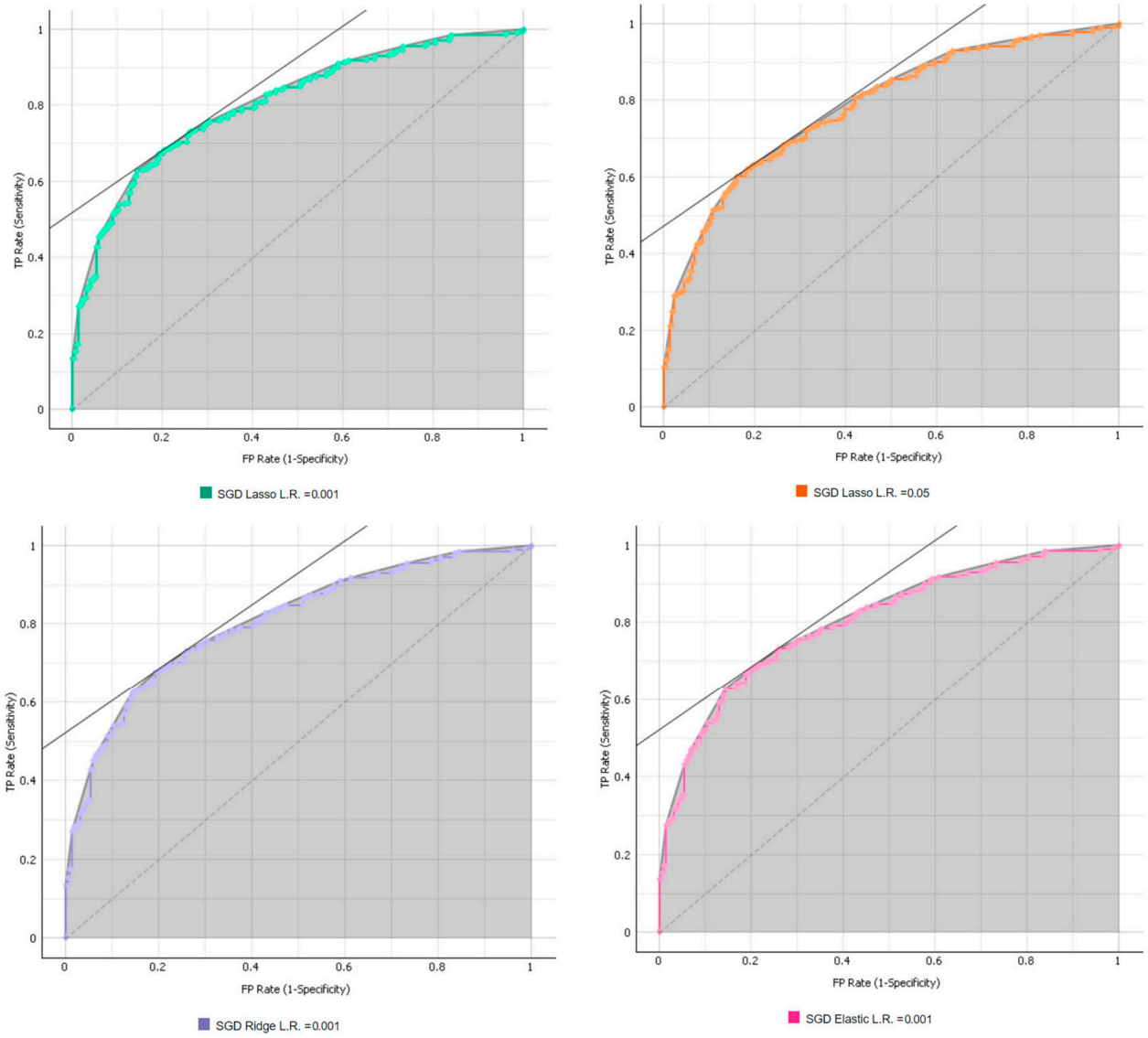

Figure A5. ROC Curve for Algorithm and Parameter Selection—SGD.

The best NN algorithm was identified in the training dataset when the number of neurons was over 15. However, in the test dataset, NN showed the best performance when the number of neurons was 30, 35, 55, and 60. The optimal number of neurons, as shown in Table A7, was 30.

Figure A6 shows the representative ROC curves for NN. The upper left graph is the ROC graph for NN with one neuron; the left lower graph is the ROC graph for NN with 50 neurons; the right upper graph is the ROC graph for NN with 30 neurons; the lower right graph is the ROC graph for NN with 100 neurons. AUC was estimated by examining the dark area under the curve. As shown in Figure A6, the ROC curves were convex, indicating

that the NN algorithms performed well in prediction. The AUC was the largest when NN was performed with 30 neurons.

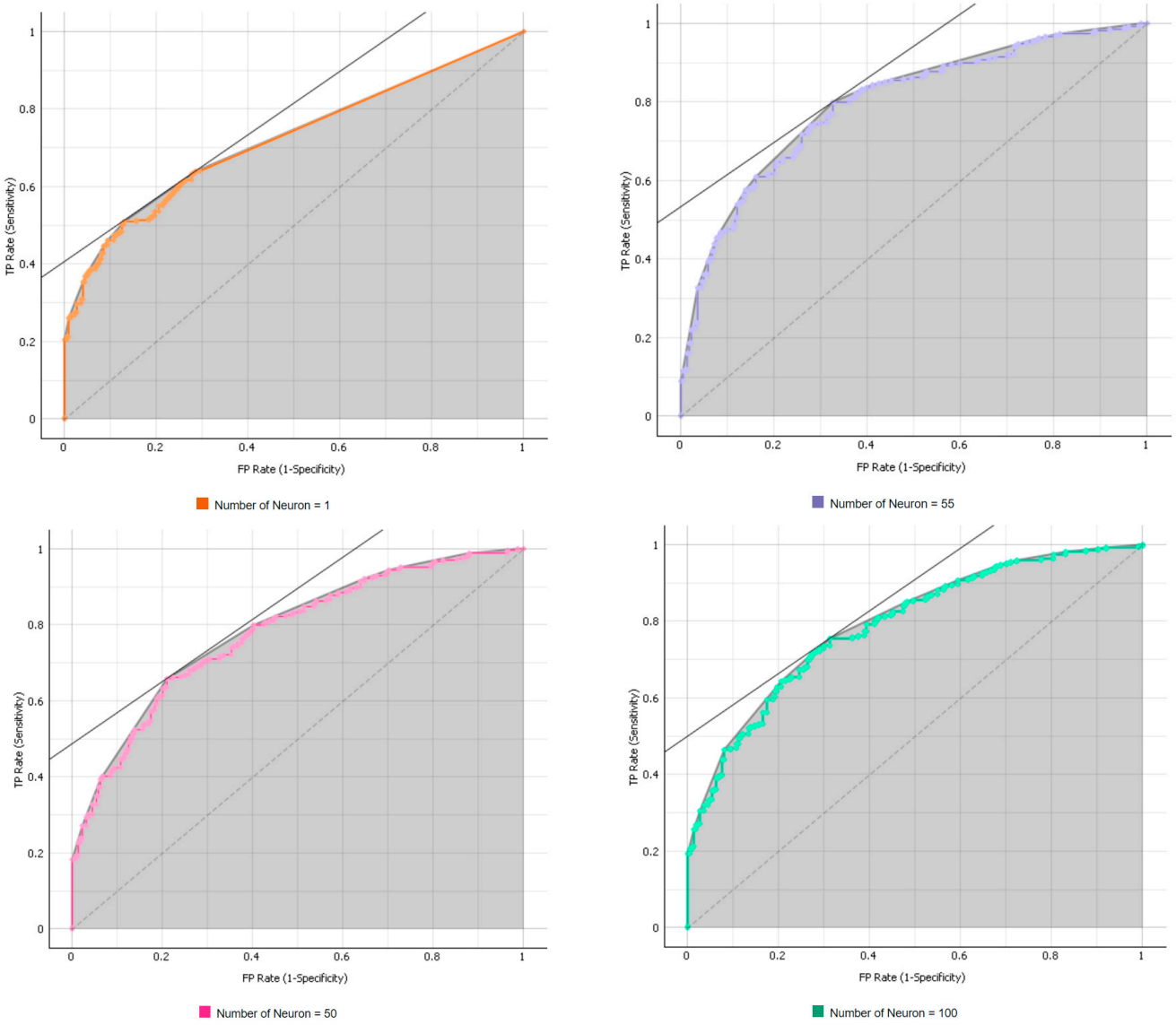

Figure A6. ROC Curve for Algorithm and Parameter Selection—NN.

Table A7. Algorithm and Parameter Selection—NN.

Number of Neuron	Training AUC	Test AUC
1	0.843	0.720
5	0.958	0.791
10	0.994	0.781
15	1.000	0.790
20	1.000	0.779
25	1.000	0.779
30	1.000	0.799
35	1.000	0.786
40	1.000	0.783

Table A7. *Cont.*

Number of Neuron	Training AUC	Test AUC
45	1.000	0.776
50	1.000	0.776
55	1.000	0.793
60	1.000	0.781
65	1.000	0.787
70	1.000	0.780
75	1.000	0.768
80	1.000	0.785
85	1.000	0.783
90	1.000	0.780
95	1.000	0.790
100	1.000	0.787

References

1. Bronfenbrenner, U. Toward an experimental ecology of human development. *Am. Psychol.* **1977**, *32*, 513–531. [CrossRef]
2. Salignac, F.; Hamilton, M.; Noone, J.; Marjolin, A.; Muir, K. Conceptualizing financial wellbeing: An ecological life-course approach. *J. Happiness Stud.* **2020**, *21*, 1581–1602. [CrossRef]
3. Despard, M.R.; Friedline, T.; Martin-West, S. Why do households lack emergency savings? The role of financial capability. *J. Fam. Econ. Issues* **2020**, *41*, 542–557. [CrossRef]
4. Gjertson, L. Emergency Saving and Household Hardship. *J. Fam. Econ. Issues* **2016**, *37*, 1–17. [CrossRef]
5. Wang, W.; Cui, Z.; Chen, R.; Wang, Y.; Zhao, X. Regression Analysis of Clustered Panel Count Data with Additive Mean Models. Statistical Papers. Advanced Online Publication. 2023. Available online: https://link.springer.com/article/10.1007/s00362-023-01511-3#citeas (accessed on 1 November 2023).
6. Heo, W. *The Demand for Life Insurance: Dynamic Ecological Systemic Theory Using Machine Learning Techniques*; Springer: Berlin/Heidelberg, Germany, 2020.
7. Luo, C.; Shen, L.; Xu, A. Modelling and estimation of system reliability under dynamic operating environments and lifetime ordering constraints. *Reliab. Eng. Syst. Saf.* **2022**, *218 Pt A*, 108136. [CrossRef]
8. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
9. Carmona, P.; Climent, F.; Momparler, A. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. *Int. Rev. Econ. Financ.* **2019**, *61*, 304–323. [CrossRef]
10. Guelman, L. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Experts Syst. Appl.* **2012**, *39*, 3659–3667. [CrossRef]
11. Heo, W.; Lee, J.M.; Park, N.; Grable, J.E. Using artificial neural network techniques to improve the description and prediction of household financial ratios. *J. Behav. Exp. Financ.* **2020**, *25*, 100273. [CrossRef]
12. Jadhav, S.; He, H.; Jenkins, K. Information gain directed genetic algorithm wrapper feature selection for credit rating. *Appl. Soft Comput.* **2018**, *69*, 541–553. [CrossRef]
13. Kalai, R.; Ramesh, R.; Sundararajan, K. Machine Learning Models for Predictive Analytics in Personal Finance. In *Modeling, Simulation and Optimization*; Das, B., Patgiri, R., Bandyopadhyay, S., Balas, V.E., Eds.; Smart Innovation, Systems and Technologies; Springer: Singapore, 2022; Volume 292.
14. Viaene, S.; Derrig, R.A.; Dedene, G. A case study of applying boosting Naïve Bayes to claim fraud diagnosis. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 612–620. [CrossRef]
15. Zhang, Y.; Haghni, A. A gradient boosting method to improve travel time predictions. *Transp. Res. Part C-Emerg. Technol.* **2015**, *58 Pt B*, 308–324. [CrossRef]
16. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361. [CrossRef]
17. Harness, N.; Diosdado, L. Household financial ratios. In *De Gruyter Handbook of Personal Finance*; Grable, J.E., Chatterjee, S., Eds.; De Gruyter: Berlin, Germany, 2022; pp. 171–188.
18. Johnson, D.P.; Widdows, R. Emergency fund levels of households. In Proceedings of the 31st Annual Conference of the American Council on Consumer Interests, Fort Worth, TX, USA, 27–30 March 1985; pp. 235–241.
19. Lytton, R.H.; Garman, E.T.; Porter, N. How to use financial ratios when advising clients. *J. Financ. Couns. Plan.* **1991**, *2*, 3–23.
20. Prather, C.G.; Hanna, S. Ratio analysis of personal financial statements: Household norms. In Proceedings of the Association for Financial Counseling and Planning Education; Edmondsson, M.E., Perch, K.L., Eds.; AFCPE: Westerville, OH, USA, 1987; pp. 80–88.
21. Greninger, S.A.; Hampton, V.L.; Kim, K.A.; Achacoso, J.A. Ratios and benchmarks for measuring the financial well-being of families and individuals. *Financ. Serv. Rev.* **1996**, *5*, 57–70. [CrossRef]

22. Bi, L.; Montalto, C.P. Emergency funds and alternative forms of saving. *Financ. Serv. Rev.* **2004**, *13*, 93–109.
23. Hanna, S.; Fan, J.X.; Change, Y.R. Optimal life cycle savings. *J. Financ. Couns. Plan.* **1995**, *6*, 1–16.
24. Cagetti, M. Wealth accumulation over the life cycle and precautionary saving? *Rev. Econ. Stat.* **2003**, *80*, 410–419. [CrossRef]
25. Kudyba, S.; Kwatinetz, M. Introduction to the big data era. In *Big Data, Mining, and Analytics*; Kudyba, S., Ed.; CRC Press and Taylor and Francis: Boca Raton, FL, USA, 2014; pp. 1–15.
26. Thompson, W. Data mining methods and the rise of big data. In *Big Data, Mining, and Analytics*; Kudyba, S., Ed.; CRC Press and Taylor and Francis: Boca Raton, FL, USA, 2014; pp. 71–101.
27. Sarker, I.H. Machine learning: Algorithms, real-World applications and research directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef]
28. Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.; Arshard, H. State-of-the-art in artificial neural network applications: A survey. *Heliyon* **2018**, *4*, e00938. [CrossRef]
29. Demsar, J.; Curk, T.; Erjavec, A.; Gorup, C.; Hočevar, T.; Milutinovič, M.; Možina, M.; Polajnar, M.; Toplak, M.; Starič, A.; et al. Orange: Data mining toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349–2353.
30. Pisner, D.A.; Schnyer, D.M. Chapter 6—Support vector machine. In *Machine Learning*; Mechelli, A., Vieira, S., Eds.; Academic Press: Cambridge, MA, USA, 2020; pp. 101–121.
31. Rudin, C.; Daubechies, I.; Schapire, R. Fin The dynamics of AdaBoost: Cyclic behavior and convergence of margins. *J. Mach. Learn. Res.* **2004**, *5*, 1557–1595.
32. Suthaharan, S. Support Vector Machine. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*; Suthaharan, S., Ed.; Springer: New York, NY, USA, 2016; pp. 207–235.
33. Meng, Y.; Li, X.; Zheng, X.; Wu, F.; Sun, X.; Zhang, T.; Li, J. Fast Nearest Neighbor Machine Translation. *arXiv* **2021**, arXiv:2105.14528.
34. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Philip, S.Y.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [CrossRef]
35. Triguero, I.; Garcia-Gil, D.; Maillo, J.; Luengo, J.; Garcia, S.; Herrera, F. Transforming big data into smart data: An insight on the use of the k-nearest neighbor algorithms to obtain quality data. *WIREs Data Min. Knowl. Discov.* **2018**, *9*, e1289. [CrossRef]
36. Fix, E.; Hodges, J.L. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev. Rev. Int. De Stat.* **1989**, *57*, 238–247. [CrossRef]
37. Singh, A.; Yadav, A.; Rana, A. K-means with three different distance metrics. *Int. J. Comput. Appl.* **2013**, *67*, 13–17. [CrossRef]
38. Östermark, R. A fuzzy vector valued KNN-algorithm for automatic outlier detection. *Appl. Soft Comput.* **2009**, *9*, 1263–1272. [CrossRef]
39. Maede, N. A comparison of the accuracy of short-term foreign exchange forecasting methods. *Int. J. Forecast.* **2002**, *18*, 67–83. [CrossRef]
40. Phongmekin, A.; Jarumaneeroj, P. Classification Models for Stock's Performance Prediction: A Case Study of Finance Sector in the Stock Exchange of Thailand. In Proceedings of the 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST), Phuket, Thailand, 4–7 July 2018; pp. 1–4.
41. Breiman, L. *Arcing the Edge*; Technical Report 486; Statistics Department, University of California at Berkeley: Berkeley, CA, USA, 1997.
42. Friedman, J.H. Greedy function approximation: A Gradient Boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
43. Sagi, O.; Rokach, L. Ensemble learning: Survey. *WIREs Data Min. Knowl. Discov.* **2017**, *8*, e1249. [CrossRef]
44. Chang, Y.; Chang, K.; Wu, G. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Appl. Soft Comput.* **2018**, *73*, 914–920. [CrossRef]
45. Liu, J.; Wu, C.; Li, Y. Improving financial distress prediction using financial network-based information and GA-based Gradient Boosting model. *Comput. Econ.* **2017**, *53*, 851–872. [CrossRef]
46. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient Boosting with Categorical Features Support. *arXiv* **2018**, arXiv:1810.11363v1. [CrossRef]
47. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
48. Hand, D.J.; Yu, K. Idiot's Bayes—Not so stupid after all? *Int. Stat. Rev.* **2001**, *69*, 385–398.
49. Lowd, D.; Domingos, P. Naïve Bayes models for probability estimation. In Proceedings of the ICML '05: Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; pp. 529–536.
50. Zhang, H. Exploring conditions for the optimality of Naïve Bayes. *Int. J. Pattern Recognit. Artif. Intell.* **2005**, *19*, 183–198. [CrossRef]
51. Yang, F. An implementation of Naïve Bayes classifier. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 12–14 December 2018; pp. 301–306.
52. Deng, Q. Detection of fraudulent financial statements based on Naïve Bayes classifier. In Proceedings of the 2010 5th International Conference on Computer Science and Education, Hefei, China, 24–27 August 2010; pp. 1032–1035.
53. Shihavuddin, A.S.M.; Ambia, M.N.; Arefin, M.M.N.; Hossain, M.; Anwar, A. Prediction of stock price analyzing the online financial news using Naïve Bayes classifier and local economic trends. In Proceedings of the 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), Chengdu, China, 20–22 August 2010; pp. V4-22–V4-26.
54. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]

55. Yu, L.; Yao, X.; Wang, S.; Lai, K.K. Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Syst. Appl.* **2011**, *38*, 15392–15399. [CrossRef]
56. Chen, F.; Li, F. Combination of feature selection approaches with SVM in credit scoring. *Expert Syst. Appl.* **2010**, *37*, 4902–4909. [CrossRef]
57. Chen, W.; Du, Y. Using neural networks and data mining techniques for the financial distress prediction model. *Expert Syst. Appl.* **2009**, *36*, 4075–4086. [CrossRef]
58. Baesens, B.; Van Gestel, T.; Viaene, S.; Stepanova, M.; Suykens, J.; Vanthienen, J. Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Oper. Res. Soc.* **2003**, *54*, 627–635. [CrossRef]
59. Yang, Y. Adaptive credit scoring with kernel learning methods. *Eur. J. Oper. Res.* **2007**, *183*, 1521–1536. [CrossRef]
60. Kim, K.; Ahn, H. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Comput. Oper. Res.* **2012**, *39*, 1800–1811. [CrossRef]
61. Chaudhuri, A.; De, K. Fuzzy support vector machine for bankruptcy prediction. *Appl. Soft Comput.* **2011**, *11*, 2472–2486. [CrossRef]
62. Chen, L.; Hsiao, H. Feature selection to diagnose a business crisis by using a real Ga-based support vector machine: An empirical study. *Expert Syst. Appl.* **2008**, *35*, 1145–1155. [CrossRef]
63. Hsieh, T.; Hsiao, H.; Yeh, W. Mining financial distress trend data using penalty guided support vector machines based on hybrid of particle swarm optimization and artificial bee colony algorithms. *Neurocomputing* **2012**, *82*, 196–206. [CrossRef]
64. Amari, S. A theory of adaptive pattern classifiers. *IEEE Trans. Electron. Comput.* **1967**, *EC-16*, 299–307. [CrossRef]
65. Amari, S. Backpropagation and stochastic gradient descent method. *Neurocomputing* **1993**, *5*, 185–196. [CrossRef]
66. Ketkar, N. Stochastic Gradient Descent. In *Deep Learning with Python*; Apress: Berkeley, CA, USA, 2017.
67. Song, S.; Chaudhuri, K.; Sarwate, A.D. Stochastic gradient descent with differentially private updates. In Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, 3–5 December 2013; pp. 245–248.
68. Newton, D.; Pasupathy, R.; Yousefian, F. Recent trends in stochastic gradient decent for machine learning and big data. In Proceedings of the 2018 Winter Simulation Conference (WSC), Gothenburg, Sweden, 9–12 December 2018; pp. 366–380.
69. Deepa, N.; Prabadevi, B.; Maddikunta, P.K.; Gadekallu, T.R.; Baker, T.; Khan, M.A.; Tariq, U. An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. *J. Supercomput.* **2020**, *77*, 1998–2017. [CrossRef]
70. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [CrossRef]
71. Matías, J.M.; Vaamonde, A.; Taboada, J.; González-Manteiga, W. Support vector machines and gradient boosting for graphical estimation of a slate deposit. *Stoch. Environ. Res. Risk Assess.* **2004**, *18*, 309–323. [CrossRef]
72. Moisen, G.G.; Freeman, E.A.; Blackard, J.A.; Frescino, T.S.; Zimmermann, N.E.; Edward, T.C., Jr. Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecol. Model.* **2006**, *199*, 176–187. [CrossRef]
73. Baum, E.B. Neural nets for economics. In *The Economy as an Evolving Complex System, Proceedings of the Evolutionary Paths of the Global Economy Workshop, Sante Fe, NM, USA, 8–18 September 1987*; Anderson, P., Arrow, K., Pindes, D., Eds.; Addison-Wesley: Reading, MA, USA, 1988; pp. 33–48.
74. Kirkos, E.; Spathis, C.; Manolopoulos, Y. Data mining techniques for the detection of fraudulent financial statement. *Expert Syst. Appl.* **2007**, *32*, 995–1003. [CrossRef]
75. Cerullo, M.J.; Cerullo, V. Using neural networks to predict financial reporting fraud: Part 1. *Comput. Fraud. Secur.* **1999**, *5*, 14–17.
76. Dorronsoro, J.R.; Ginel Fin Sgnchez, C.; Cruz, C.S. Neural fraud detection in credit card operations. *IEEE Trans. Neural Netw.* **1997**, *8*, 827–834. [CrossRef]
77. Chauhan, N.; Ravi, V.; Chandra, D.K. Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks. *Expert Syst. Appl.* **2009**, *36*, 7659–7665. [CrossRef]
78. Iturriaga, F.J.L.; Sanz, I.P. Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks. *Expert Syst. Appl.* **2015**, *42*, 2857–2869. [CrossRef]
79. Menard, S. *Applied Logistic Regression Analysis*, 2nd ed.; Sage Publications: Thousand Oaks, CA, USA, 2002.
80. Arcuri, A.; Fraser, G. Parameter tuning or default values? An empirical investigation in search-based software engineering. *Empir. Softw. Eng.* **2013**, *18*, 594–623. [CrossRef]
81. Joseph, V.R. Optimal ratio for data splitting. *Stat. Anal. Data Min.* **2022**, *15*, 531–538. [CrossRef]
82. Afendras, G.; Markatou, M. Optimality of training/test size and resampling effectiveness in cross-validation. *J. Stat. Plan. Inference* **2019**, *199*, 286–301. [CrossRef]
83. Picard, R.R.; Berk, K.N. Data Splitting. *Am. Stat.* **1990**, *44*, 140–147.
84. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
85. Kira, K.; Rendell, L.A. A practical approach to feature selection. In *Machine Learning: Proceedings of International Conference (ICML'92)*; Sleeman, D., Edwards, P., Eds.; Morgan Kaufmann: Burlington, MA, USA, 1992; pp. 249–256.
86. Kononenko, I. Estimating attributes: Analysis and extensions of Relief. In *Machine Learning: ECML-94*; De Raedt, L., Bergadano, F., Eds.; Springer: Berlin/Heidelberg, Germany, 1994; pp. 171–182.
87. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69. [CrossRef]
88. Heo, W.; Cho, S.; Lee, P. APR Financial Stress Scale: Development and Validation of a Multidimensional Measurement. *J. Financ. Ther.* **2020**, *11*, 2. [CrossRef]

89. Xiao, J.J.; Ahn, S.Y.; Serido, J.; Shim, S. Earlier financial literacy and later financial behavior of college students. *Int. J. Consum. Stud.* **2014**, *38*, 593–601. [CrossRef]
90. Lusardi, A. Financial literacy and the need for financial education: Evidence and implications. *Swiss J. Econ. Stat.* **2019**, *155*, 1. [CrossRef]
91. Grable, J.E.; Lytton, R.H. Financial risk tolerance revisited: The development of a risk assessment instrument. *Financ. Serv. Rev.* **1999**, *8*, 163–191. [CrossRef]
92. Loibl, C.; Hira, T.K. Self-directed financial learning and financial satisfaction. *J. Financ. Couns. Plan.* **2005**, *16*, 11–22.
93. Lown, J.M. Development and validation of a financial self-efficacy scale. *J. Financ. Couns. Plan.* **2011**, *22*, 54–63.
94. Perry, V.G.; Morris, M.D. Who is in control? The role of self-perception, knowledge, and income in explaining consumer financial Behavior. *J. Consum. Aff.* **2005**, *39*, 299–313. [CrossRef]
95. Diener, E.; Emmons, R.A.; Larsen, R.J.; Griffin, S. The satisfaction with life scale. *J. Personal. Assess.* **1985**, *49*, 71–75. [CrossRef] [PubMed]
96. Rosenberg, M. *Society and the Adolescent Self-Image*; Princeton University Press: Princeton, NJ, USA, 1965.
97. Hellgren, J.; Sverke, M.; Isaksson, K. A two-dimensional approach to job insecurity: Consequences for employee attitudes and well-being. *Eur. J. Work. Organ. Psychol.* **1999**, *8*, 179–195. [CrossRef]