Retos
Revista de Ciencias de Administración y Economía

october 2023-march 2024
pp.267-283
https://doi.org/10.17163/ret.n26.2023.06

Gaussian process regression´s hyperparameters optimization to predict financial distress

Optimización de hiperparámetros de regresión del proceso gaussiano para predecir problemas financieros

Amine Sabek
Professor at the University of Tamanrasset, Argelia
sabek.amine@univ-tam.dz
https://orcid.org/0000-0002-6970-4183

Jakub Horák
Professor at the Institute of technology and Business České Budějovice, Czech Republic
horak@mail.vstecb.cz
https://orcid.org/0000-0001-6364-9745

Received on: 07/06/23 **Revised on:** : 06/07/23 **Approved on:** 03/08/23 **Published on:** 1/10/23

Abstract: predicting financial distress has become one of the most important topics of the hour that has swept the accounting and financial field due to its significant correlation with the development of science and technology. The main objective of this paper is to predict financial distress based on the Gaussian Process Regression (GPR) and then compare the results of this model with the results of other deep learning models (SVM, LR, LD, DT, KNN). The analysis is based on a dataset of 352 companies extracted from the Kaggle database. As for predictors, 83 financial ratios were used. The study concluded that the use of GPR achieves very relevant results. Furthermore, it outperformed the rest of the deep learning models and achieved first place equally with the SVM model with a classification accuracy of 81%. The results contribute to the maintenance of the integrated system and the prosperity of the country's economy, the prediction of the financial distress of companies and thus the potential prevention of disruption of the given system.

Keywords: financial distress, Gaussian process regression, deep learning, investment financing, financial risk prediction, Gaussian regression, financial ratios, deep learning models.

Resumen: la predicción de las dificultades financieras se ha convertido en uno de los temas más importantes en el área contable y financiera debido a su correlación significativa con el desarrollo de la ciencia y la tecnología. El objetivo principal de este trabajo es predecir la dificultad financiera con base en la Regresión de Procesos Gaussianos (GPR) y luego comparar los resultados de este modelo con los resultados de otros modelos de aprendizaje profundo (SVM, LR, LD, DT, KNN). El análisis se basa en un conjunto de datos de 352 empresas extraídos de la base de datos de Kaggle. En cuanto a los predictores, se utilizaron 83 ratios financieros. El estudio concluyó que el uso de la GPR logra resultados muy relevantes. Además, superó al resto de los modelos de aprendizaje profundo y logró el primer lugar por igual con el modelo SVM con una precisión de clasificación del 81 %. Los resultados contribuyen al mantenimiento del sistema integrado y a la prosperidad de la economía del país, a la predicción de las dificultades financieras de las empresas y, por lo tanto, a la posible prevención de perturbaciones del sistema en cuestión.

Palabras clave: dificultades financieras, regresión del proceso gaussiano, aprendizaje profundo, financiamiento de inversiones, predicción del riesgo financiero, regresión gaussiana, coeficientes financieros, modelos de aprendizaje profundo.

Suggested citation: Sabek, A. and Horák, J. (2023). Gaussian process regression´s hyperparameters optimization to predict financial distress. *Retos Revista de Ciencias de la Administración y Economía, 13*(26), 267-283. https://doi.org/10.17163/ret.n26.2023.06

Introduction

Financial knowledge is essential for a given business entity. The financial health expresses the good financial situation of the company. A company is financially healthy if it guarantees the invested funds (yield, profitability), is financially stable, is not limited in its decision-making by other entities (indebtedness, financial structure), can pay its obligations and thereby guarantee the existence and appreciation of the invested funds (Gavurova *et al.*, 2020; Krulicky and Horak, 2021).

On the other hand, a financial distress can be defined as a situation in which the cash flow of a company is restricted for some reason. This restriction may be temporary if directors have the opportunity and ability to conduct corrective procedures (Liew *et al.*, 2023). Horak *et al.* (2020) mention similar characteristics of financial distress, and define it as a state in which the financial health of the company is significantly weakened. The authors also say that in case of financial distress, the company finds it challenging to draw up a payment schedule and pay its financial obligations on time within the pre-agreed maturity dates, thus exposing the company to the potential risk of legal enforcement. In such a situation, the company shows serious liquidity problems (ability to pay), and the solution amounts to significant changes in the company's operational activities and funding method (Vochozka *et al.*, 2020). The financial distress is also the final stage of organizational decline before bankruptcy. Therefore, the financial distress differs from bankruptcy as it prescribes a time when the borrower cannot pay the debts to the creditor (Hantono, 2019). The exact definition of financial distress has not yet been determined, but economic difficulties are known to have varying degrees. Mild financial distress refers to the temporary distress in cash flow and concepts such as insolvency, default, etc. The most dangerous of these degrees is bankruptcy or business failure (Shi and Li, 2019).

The importance of predicting financial distress has evolved gradually since almost half a century ago when this contemporary phenomenon appeared with the development of business establishments, where the sudden collapse of many companies was incomprehensible. Kliestik *et al.* (2018) claim that several scientific works have devoted to the issue of predicting financial distress, in order to predict the company's failure and classify the company according to its financial health. For this purpose, several methods have been used that differ in their assumptions and complexity. Anticipating financial distress before they occur, however, remains one of the solutions that have proven effective in preventing them. Initially, statistical techniques were used to build models with predictive capacity, and model building was associated with the development of science and technology. The more science develops, the more scientists and researchers who design more complex, accurate, and quality models that fill the gaps in earlier studies. The development of science has led to a revolution in the field of forecasting, where artificial intelligence techniques have been exploited in this field, achieving impressive results that are almost perfect (Bonello *et al.*, 2018). Artificial intelligence techniques to predict financial problems became common in the 1990s, with the development of computer techniques (Paule-Vianez, 2019). Deep learning has emerged and is progressively evolving into a robust technique for various uses, and has helped solve various problems in the economy and business, such as speech recognition, natural language processing, automatic driving, computer vision, prediction of financial difficulties and credit assessment (Qu *et al.*, 2019).

Several scientific studies on the subject of financial distress and forecasting of bankruptcies have proposed various predictive models for this purpose. Most published studies used data from a year before distress. Only some studies used data from 2-3 years before distress. The results showed that the data corresponding to two years before distress reduced the ability of the model to predict financial distress (Fernández-Gámez *et al.*, 2016), with accuracies of 72.0% and 95.5%, 86.2%, 100% using genetic algorithms and neural network one, two and three years before distress. Some authors compared the classification accuracy of forecasting models based on Polish industrial companies. Through R programs, the research tested neural networks, logistic regression, supporting vector machines, classification trees, k-NN

Print ISSN: 1390-6291; electronic ISSN: 1390-8618

algorithms, bagging, random forests, discriminatory analysis, empowerment and naive Bayes (Costa *et al*., 2022). Other authors have studied several intelligent and statistical models, such as particle swarms optimization integrated into semiconductor virtual machines (SVMs), decision trees, linear discriminant and genetic algorithms, using logistic regression of SVMs, self-organizing maps and quantification of learning vectors. The results show that statistical techniques are more suitable for large data sets, and smart techniques are more suitable for smaller data sets (Zhou *et al*., 2019). This improved method combines features of fuzzy sets, and machine learning can be compared to probabilistic neural networks in terms of clustering performance. The aim of this study is to predict the decline using the GP method and its subsequent comparison with logistic regression machines and support vectors. The research is based on accurate data on bankruptcy, and concluded that Gaussian processes outperform other methods in predicting bankruptcy with high accuracy (Liu *et al*., 2023).

This study aims to establish a general idea of the advantages that can be applied to the various actors, both academic and professional. The economy of a State functions as an interconnected system that encompasses many factors that contribute to the establishment of a strong and prosperous society. If any of these factors fail to meet their targets, the entire system will fail. Since economic enterprises play a fundamental role in a country's economy, it is necessary to ensure their continuity by all possible means. As a result, the importance of predicting financial distress arises as a method based on the advancement of statistical and intelligent techniques that help companies avoid bankruptcy and the cessation of their operations.

Our paper stands out among the limited number of scientific publications that address this topic, distinguishing itself by its focus on predicting financial distress using the GPR model, therefore, preliminary tests were performed on the GPR model. Our main goal is to improve academic research and make significant contributions to its advancement. For this research, two research questions were defined: Is the GPR model suitable to predict financial distress? Did the GPR model hold up against the logistic regression model to predict financial distress?

The article is structured as follows. Section 1 presents a brief review of the literature, section 2 provides information on the investigation procedure, data and variables, section 3 presents the results obtained, section 4 analyzes the results obtained and provides an overall summary of the results of the research, including the proposed recommendations.

Methodology

Data and Variables

This dataset encompasses two distinct types of variables. First, there is the independent variable X, which is a quantitative variable encompassing a range of 83 financial ratios. Unfortunately, the specific names to these ratios were not explicitly provided; instead, they were named X1, X2, ..., X83. Although this lack of accurate identification is an inconvenience, it was chosen to use these data because of its alignment with the main objective of the study, which involves assessing the predictive capacity of the model for financial difficulties after the optimization of its hyperparameters. The identification of the set of financial ratios that exert the greatest influence on the dependent variable served as a secondary objective, especially after the application of the PCA technique to enhance data quality.

The second type of variable corresponds to the dependent variable, named Y, which is a qualitative variable representing the model's outputs and covers two fundamental scenarios: financial distress, denoted as 0, and non-financial distress, named as 1. These data provide an accurate description of the actual circumstances of all financial cases, taking into account the importance of the indicators (83). Consequently, this data s*et al*lows us to effectively train the model and assess its predictive capacity.

We are relied on a ready-made data set that includes data from 352 companies extracted from the Kaggle database. We divided these data into a

training sample and a test sample, where the training sample contained the data of 187 companies for different years, and the number of financial cases (fiscal years) reached 2001 financial cases divided into 896 cases of financial distress and 1105 cases of non-financial distress, while the test sample included the data of 165 companies for a period of four years, where data for the rest of years were excluded. The number of financial cases (fiscal years) in the test sample reached 660 financial cases, divided into 351 cases of financial distress and 309 cases of non-financial distress. As for the predictors used in this study, 83 financial ratios are included, representing a considerable number of independent variables, which is desirable, since it will help us extract the most influential components in dependent variability after activating the Principal Component Analysis (PCA) technique. The key variables on which this study is based are shown in Table 1.

Table 1
Main study variables

Variable	Function
Gaussian Process Regression	Non-parametric Kernel-based probabilistic model. There is a latent variable f(xi) introduced for each observation $_{xi}$, which makes the GPR model non-parametric. In vector form, this model is equivalent to $P(y\backslash f,X) \sim N(y\backslash H\beta + f, \sigma 2I)$, Where $X=\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$, $y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $H=\begin{pmatrix} h(x_1^T) \\ h(x_2^T) \\ \vdots \\ h(x_n^T) \end{pmatrix}$, $f=\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}$ The combined distribution of the latent variables f($x_{1)}$ $f(x2)$, ..., f(xn) in the GPR model is as follows: $P(f\backslash X) \sim N(f\backslash 0, k(X, X))$, close to a linear regression model, where K(X,X) looks like this: $K(X,X)=\begin{matrix} k(x_1,x_1) & k(x_1,x_2) & k(x_1,x_n) \\ k(x_2,x_1), & k(x_2,x_2), & k(x_2,x_n) \\ k(x_n,x_1) & k(x_n,x_2) & k(x_n,x_n) \end{matrix}$
Kernel Function	In supervised learning, points with similar predictive values $_{xi}$, are expected to have $_{yi}$ near response (target) values $_{yi}$. In Gaussian processes, the covariance function expresses this similarity. This specifies the covariance between the two latent variables f(xi) and f(xj), where xi and $_{xj}$ are vectors of $_{xd}$-by-1. In other words, it determines how the response at one point xi is affected by the responses at other points xj, i ≠ j, i = 1, 2, ..., n. The covariance function $k(_{xi,xxj})$ can be defined by several Kernel functions. It can be analyzed in terms of the Kernel parameters in the vector θ. Therefore, it is possible to express the covariance function as $k(_{xi,xj}\backslash\theta)$.
Optimizing Hyperparameters	Changing the functions of the model to improve its performance.
Analysis of principal components	The analysis of principal components is a quantitatively rigorous method for achieving this simplification. The method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All principal components are orthogonal to each other, so there is no redundant information. The main components as a whole constitute an orthogonal basis for the space of the data.

Print ISSN: 1390-6291; electronic ISSN: 1390-8618

Variable	Function
Rational quadratic Kernel	This covariance function is defined by $$k(xi,xj\backslash\theta)=\sigma_f^2\left(1+\frac{r^2}{2\alpha\sigma_l^2}\right)^{-\alpha}$$ when $_l$ is the characteristic length scale, α is a scale mixing parameter with a positive value; and $$r=\sqrt{(x_i-\boldsymbol{x_j})^T(x_i-\boldsymbol{x_j})}$$
Square exponential Kernel	This is one of the most commonly used covariance functions. The square exponential function of the kernel is defined as $$k(xi,xj\backslash\theta)=\sigma_f^2\exp\left[\frac{1}{2}\frac{(x_i-\boldsymbol{x_j})^T(x_i-\boldsymbol{x_j})}{\sigma_l^2}\right];$$ where $_{\alpha l}$ is the characteristic length scale and f is the standard deviation of the signal.
Matern 5/2 Kernel	This covariance function is defined by $$k(xi,xj)=\sigma_f^2\left(1+\frac{\sqrt{5r}}{\sigma_l}+\frac{5r^2}{3\sigma_l^2}\right)exp\left(-\frac{\sqrt{5r}}{\sigma_l}\right)$$ where $$r=\sqrt{(x_i-\boldsymbol{x_j})^T(x_i-\boldsymbol{x_j})}$$
Exponential Kernel	This covariance function is defined by $$k(xi,xj\backslash\theta)=\sigma_f^2\exp\left(-\frac{r}{\sigma_l}\right)$$ where $_{\alpha l}$ is the characteristic length scale; and $$r=\sqrt{(x_i-\boldsymbol{x_j})^T(x_i-\boldsymbol{x_j})}$$
Matern 3/2 Kernel	This covariance function is defined by $$k(xi,xj\backslash\theta)=\sigma_f^2\left(1+\frac{\sqrt{3r}}{\sigma_l}\right)\text{expexp}\left(-\frac{\sqrt{3r}}{\sigma_l}\right)$$ where $$r=\sqrt{(x_i-\boldsymbol{x_j})^T(x_i-\boldsymbol{x_j})}$$

Methods

In this study, a descriptive methodology was used in the theoretical section, using accredited sources and peer-reviewed indexed academic journals. On the other hand, the section applied had a comparative approach using an analytical method. Two primary inquiries were made, and after extracting the results related to the improvement of the quality of the GPR model, a comparison with deep learning models was made. The findings were thoroughly discussed and the main research was addressed.

Matlab was used to build the model, optimizing its hyperparameters and being able to show the resulting graphs. Excel was used to calculate the error measures (MSE, RMSE, MAE), as well as to determine the elements of the confusion matrix (Sensitivity, Specificity, Accuracy). The SPSS was used to perform the R^2 statistical test. The statistical evaluation of the model was performed using the measure R^2, widely considered as one of the most significant statistical tests due to its ability to assess the correlation between real and predicted values. No additional statistical tests were performed, with the exception of the R^2 test, as the researcher believed that the R^2 test adequately captured the statistical significance of the model. In addition, it should be noted that PCA was deemed ineffective. The model was subjected to a mathematical evaluation using several significant mathematical measures, including MAE, RMSE and MSE, to quantify the error of the model. Additionally, the evaluation involved the examination of the matrix of confusion and its associated metrics, such as Sensitivity, Specificity, Accuracy. These measures were used to compare the performance of the GPR model with that of deep learning models.

The temporal and spatial scope of this study is not available, and as noted above, these data were extracted from the Kaggle database, and are available at the following link: https://bit.ly/3DZxGr1. Unfortunately, despite the importance of these two aspects, the data available do not provide specific information on time and space coverage. However, due to the need for valuable and meaningful results and the absence of superior alternatives, we have chosen to rely on this dataset. The dataset is remarkable, as its owner reports that it possesses the following attributes: well-documented, well-maintained, clean, and original data. In addition, it covers a wide time range, although the exact period is not specified. This scope allows us to evaluate the predictive capacity of the models in forecasting financial distress four years prior to its occurrence.

In the first phase of this study, five types of GPR models will be formulated, each one distinguished by the type of Kernel function used. Subsequently, these models will be subjected to training using the training sample provided, after which a comparative analysis will be performed to identify the most optimal model that exhibits the minimum error value. The selected model then goes to the second stage for testing. Additionally, after evaluating this model, additional types of models will be formulated using the same Kernel function that achieved the best results in the previous phase, but varying in terms of the base function used. Again, a selection process will be carried out to determine the most optimal model, which will then move to the final phase that requires comparison between the extracted models and the commonly used machine learning models.

By using Gaussian processes, a good framework for probability regression can be provided (Yang *et al.*, 2023). The Gaussian process method has recently reborn due to the advent of artificial intelligence and Kernel-based machine learning. These models provide various uses in various areas of research and a complete nonlinear Bayesian method (Antunes *et al.*, 2017). GPR is a non-parametric model that depends on the probability distribution of Gauss and is defined as a set of random variables. Each finite GP number of this random variable has a common Gaussian distribution. Thus, GP is fully specified by the second-order statistic:

$$f(x) \sim GP(m(x), k(x, x')) \quad (1)$$

Where $m(x)$ and $k(x,x')$ are the covariance and mean functions of a real process $f(x)$, respectively (Ferkousl *et al.*, 2021). It only defines the covariance and mean functions to simplify a function from

Print ISSN: 1390-6291; electronic ISSN: 1390-8618

a Gaussian process. The k covariance function models the articular variability of the random variables from the Gaussian process, and returns the modeled covariance between the pair of inputs (Herfurth, 2020). The Gaussian process is a robust non-parametric method with precise uncertainty models, mainly used in classification and regression issues. It is not parametric because the Gaussian process tries to infer how all measured data are correlated rather than adjusting the parameters of the chosen base functions (Wang *et al.*, 2023). A Gaussian process is an operational probabilistic regression method, originally pioneering in statistics and geophysics, that has since found a strong user base in machine learning. A Gaussian process, considered a probabilistic regression technique, takes a kernel and dataset as input and gives the distribution of a function as output (Asante-Okyere *et al.*, 2018).

GPR can be considered as a generalization of the more standard Bayesian linear regression, and similarly, the classification of the Gaussian process can be considered a generalization of the logistic regression. The activation of the logistic function was given by $a = wT \varphi(x)$. Therefore, Gaussian processes can be allowed not to linearize the function by directly manipulating the function space (Hamoudi *et al.*, 2023). Therefore, we can replace the linear model $wT \varphi(x(n))$ with a Gaussian process f considering the set of latent variables for $n \in \{1, N\}$. In addition, we are interested in the probability of membership of $\pi(x^\star) = p(y = 1 \mid x^\star) = \sigma(f(x^\star))$ given an observation $x^\star$. The inference process is similar to the previous one, so the distribution of $x^\star$ is calculated as:

$$p(f^\star \mid D) = \int p(f^\star \mid D,f)\, p(f \mid D)\, \partial f \qquad (2)$$

Where $p(f \mid D) \propto p(D \mid f)\, p(f)$ is the posterior obtained by applying the Bayes rule (Taki *et al.*, 2018).

Given the function of covariance, making predictions for new test points is simple, because it is about manipulating algebraic matrices. However, in procedural application, it may be necessary to recognize which covariance function to use. Of course, the reliability of the regression depends on how well the parameters required by the chosen covariance function were selected (Wang *et al.*, 2023).

Results

In this section, we will present the results obtained through experimentation and discuss these results clearly. Having organized and distributed the data into a training sample and test sample, we will proceed to build and develop multiple models to assess its ability to predict financial distress. However, first, a comprehensive review of the data will be carried out. Box diagrams are used to illustrate the data for several reasons. First, box diagrams provide valuable information about data dispersion or variability. Second, they provide reliability of the distribution of securities. Third, they help identify regions where sample values are most densely clustered or scarce. Due to the large number of independent variables, namely 83, it is not practical to create a separate cash flow chart for each variable. Therefore, we will selectively show the box diagram for a specific set of variables, namely X1, X2, X6, X24, X30 and X81, chosen at random only for illustrative purposes. Figure 1 shows outliers, represented in red, that are seen in two areas of the figure, either above the maximum value after outliers are excluded or below the minimum value after outliers are excluded.

Figure 1
Box of Whiskers Plot

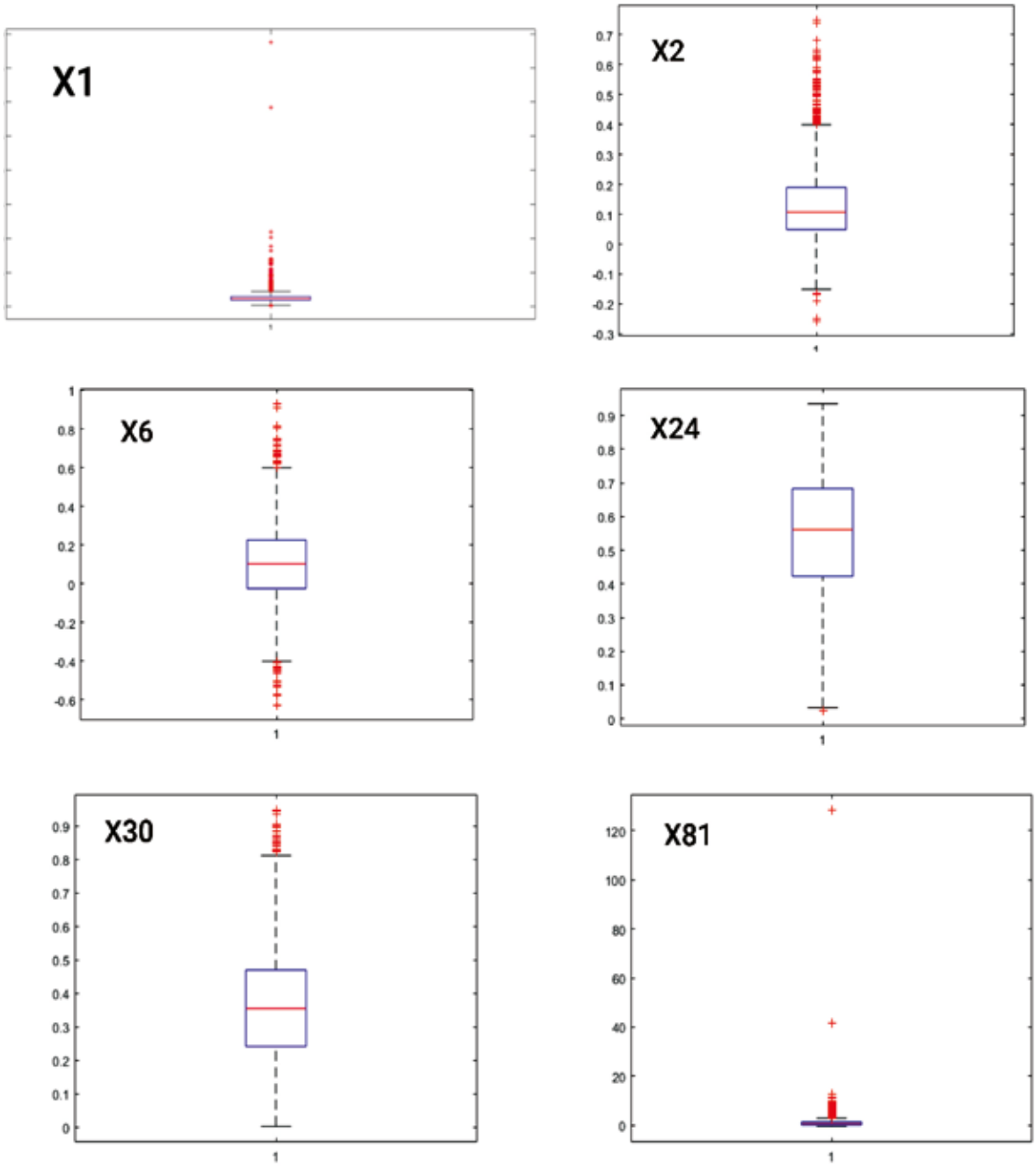


Table 2 illustrates the design features of the GPR models and provides a clear overview of all the details, as detail below.

Print ISSN: 1390-6291; electronic ISSN: 1390-8618

Table 2
Design Process Variables

Kernel Function	Base Function	Kernel Sigma	Isotropic Kernel	Standardized	Optimize Numeric Parameters
Rational quadratic	Constant	0.3517075	True	True	True
Exponential Square	Constant	0.3517075	True	True	True
Matern 5/2	Constant	0.3517075	True	True	True
Exponential	Constant	0.3517075	True	True	True
Matern 3/2	Constant	0.3517075	True	True	True

Table 2 presents the main objective of the research in the design of diverse GPR models in order to compare their outcomes and identify the most optimal model. It is important to note that throughout the design phase, all parameters remained fixed and were not affected by variations in the Kernel function type. In addition, several non-essential fields, such as prediction speed and training time, were omitted from the analysis as they hold lesser significance. As shown in Table 2, during the first phase, the PCA feature was used to extract the main components and reduce the number of predictors, due to the substantial incorporation of financial ratios. By using this widely recognized and indispensable technique, we can effectively eliminate variables that do not contribute to research objectives and hinder the attainment of accurate predictions regarding financial distress. The results obtained from the training of GPR models after the activation of the PCA technique are presented in Table 3.

Table 3
Training results using the PCA technique

Kernel Function	RMSE	MSE	MAE	R^2
Rational quadratic	0.497	0.247	0.494	0.00
Exponential Square	0.497	0.247	0.494	0.00
Matern 5/2	0.497	0.247	0.494	0.00
Exponential	0.492	0.242	0.483	0.02
Matern 3/2	0.497	0.247	0.494	0.00

Root mean square error (RMSE), mean square error (MSE) and mean absolute error (MAE) were used to measure the error value. The coefficient of determination, denoted as R^2, was used as a statistical metric to evaluate the quality of the model and to understand the correlation between the independent variables and the dependent variable, as well as the correlation between the observed values and the predicted values. When analyzing Table 3, it was observed that the results obtained could have been more satisfactory. Surprisingly, upon examining Table 3, it was observed that the obtained results could have been more satisfactory. However, these outcomes are inadequate for progressing to the second phase, namely "testing." In this phase, it was noticed that the measures of prediction accuracy were excessively inflated and almost identical across all models. Additionally, the R^2 values were nearly zero for all models, indicating a lack of correlation between the predictors and the dependent variable, rendering the models statistically insignificant. Consequently, this suggests the possibility of an imbalance resulting from the utilization of the Principal Component Analysis (PCA) technique. This observation is perplexing and contradictory, as the PCA technique typically contributes to reducing error and enhancing prediction quality. However, this expected improvement is not evident in this case. Therefore, it is necessary to

investigate the causes behind the inflated error measures and the absence of the coefficient of determination. It is proposed to disable the PCA technique and assess whether the results will exhibit improvement or further deterioration. Subsequently, in Table 4, the PCA technique is disabled, yielding the following set of results.

Table 4
Training results without using the PCA technique

Kernel function	RMSE	MSE	MAE	R^2
Rational quadratic	0,369	0,136	0,297	0,45
Exponential square	0,372	0,138	0,303	0,44
Matern 5/2	0,371	0,138	0,299	0,44
Exponential	0,370	0,137	0,295	0,45
Matern 3/2	0,371	0,137	0,298	0,44

By analyzing Table 4, a noticeable decline in the values of prediction accuracy measures becomes apparent, suggesting a decrease in error rates. This signifies an enhancement in the prediction quality of the models, which is further corroborated by the substantial increase in the values of R^2. However, it is important to note that these values did not approach 1 but remained considerably distant from zero. Consequently, the models have achieved statistical significance and can effectively elucidate the relationship between the predictors and the dependent variable with a correlation coefficient of 0.444. Hence, we can infer that the employment of Principal Component Analysis (PCA) primarily contributed to the substandard performance of the models. Upon comparing the prediction accuracy measures, it is evident that the initial model employing the Rational Quadratic Kernel function exhibits lower error values compared to the other models, as well as higher R2 values. Additionally, this model attains the highest level of statistical significance. Consequently, we will disregard the remaining models and opt to employ this particular model for testing in the subsequent phase. The test results of the Rational Quadratic model, based on the same aforementioned measures, are presented in Table 5.

Table 5
Testing results (overall)

Kernel Function	RMSE	MSE	MAE	R^2	Accuracy
Rational quadratic	0,380	0,144	0,318	0,42	0,80

Table 5 shows the results after testing the Rational Quadratic model using the test sample. We note that the prediction accuracy measures increased compared to the training phase, which is expected. On the other hand, it is positive because the error values increased only slightly, and this indicates that the model was able to build the appropriate formula that serves the objective of the study, and this can be confirmed by the prediction accuracy rate of 80%, which is a very appropriate rate and reflects the strength of the model in predicting financial distress. In order to further clarify the results of the model test, we will rely on Figure 2 and Table 6 to provide more detailed information.

Print ISSN: 1390-6291; electronic ISSN: 1390-8618

Figure 2
Predicted Vs. Actual Plot (RQ-Constant)

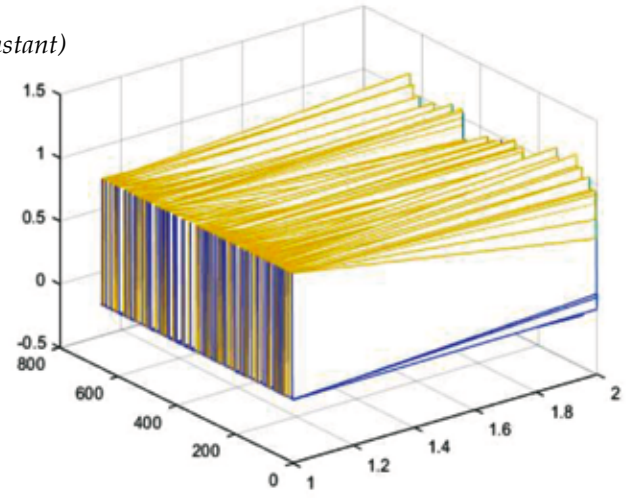


Table 6
Testing results (particular)

Kernel Function	Year	RMSE	MSE	MAE	R^2
Rational quadratic	N-1	0206	0042	0087	0,31
	N-2	0183	0033	0075	0,45
	N-3	0174	0,0305	0072	0,52
	N-4	0194	0037	0082	0,39

The model's accuracy in predicting financial distress was assessed at different time points: one year before distress occurrence, two years ago, three years ago, and four years ago. It is worth noting that the error values were highest in the first year, accompanied by a clear decrease in the R^2 value. This observation is intriguing since the classification model's performance in the initial year was expected to be superior to subsequent years, and then it begins to decline gradually. However, the opposite was correct, as the quality of prediction improved the farther away the possibility of distress occurrence. Therefore, it can be said that the model is promising because it achieved relevant results, and therefore we will optimize the model's hyperparameters to improve the results. Matlab allows us to make several modifications in the design phase of the model and before training it. Perhaps an essential feature that can be modified is related to the primary function because we have made several other modifications. However, they did not achieve appropriate results, so it is unnecessary to comment on these modifications. As shown in Table 7, the program offers three types of Basis functions, enabling the construction of three new GPR models based on these functions. However, only two new models will be created, as the Rational Quadratic model utilizing the Constant Basis function has already been constructed in the previous phase.

Table 7
GPR-RQ Hyperparameters Optimization (Training results)

Base Function	RMSE	MSE	MAE	R^2
Constant	0,369	0,136	0,297	0,45
Zero	0,368	0,136	0,296	0,45
Linear	267,13	71356	11,273	-287973

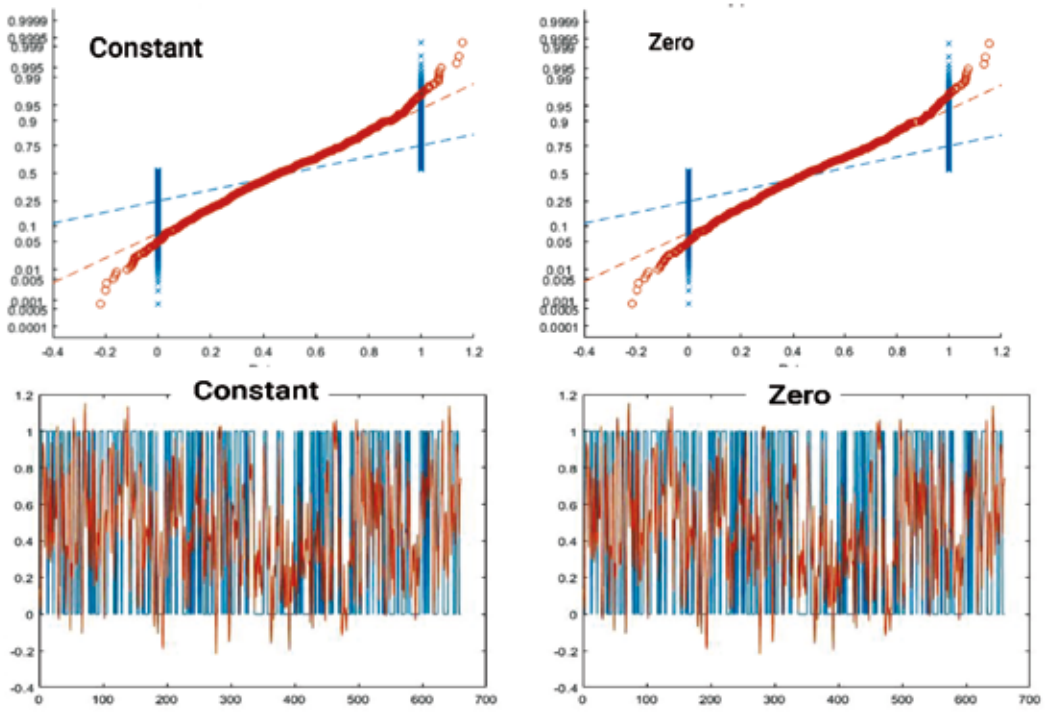
Training results of the linear relationship model were unsatisfactory and are therefore omitted. Based on the training results of the remaining two models, it is observed that both the error values and the R^2 values show convergence, although the Zero model has shown a slightly higher performance. These findings provide impetus to proceed to the testing phase and conduct a comparative analysis of the two models, as the training results have indicated the potential for enhancing the accuracy of the Rational Quadratic model. The results outlined in Table 8 present the following outcomes.

Table 8
Testing results (overall)

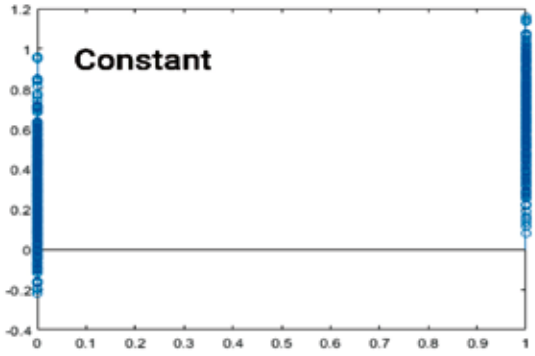
Basis Function	RMSE	MSE	MAE	R^2	Sensitivity	Specificity	Accuracy
Constant	0.380	0.144	0.318	0,42	0,82	0,78	0,80
Zero	0377	0142	0315	0,43	0,83	0,79	0,81

It should be noted that the RQ-Zero model showed superior performance compared to the RQ-Constant model for all metrics presented in Table 7. Thus, results have improved, albeit marginally. For a more complete view of the test results for both models, we will use Figure 3 and Table 9 to present more complex and detailed information. We present the Constant-GPR and Zero-GPR model figures, because the results of these two models were valuable compared to previous models. We hope to clarify the difference between the two models through the residual graph, but as observed, Figure 3 does not show a significant difference between the two models due to the convergence of the results.

Figure 3
Residual Plot



Print ISSN: 1390-6291; electronic ISSN: 1390-8618



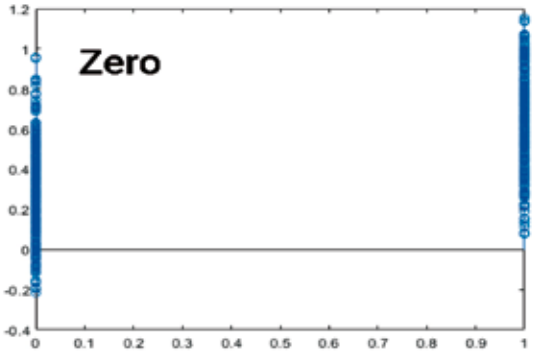


Table 9
Testing results (particular)

Función de base	Año	RMSE	MSE	MAE	R^2
Constante	N-1	0206	0042	0087	0,31
	N-2	0183	0033	0075	0,45
	N-3	0174	0,0305	0072	0,52
	N-4	0194	0037	0082	0,39
Cero	N-1	0204	0042	0086	0,32
	N-2	0182	0033	0075	0,45
	N-3	0172	0029	0072	0,53
	N-4	0193	0037	0081	0,40

Based on the data presented in Table 9, all available evidence suggests the superiority of the RQ-Zero model. It is worth noting that a similar problem found in the first model also occurred in the RQ-Zero model, where the error values were higher in the first year. This trend can also be observed in the R^2 value, as the classification capacity of the model was expected to be higher in the first year and to decrease gradually in subsequent years. The opposite was observed in this case. In the final stage, once the optimal model was identified from the GPR models, we proceeded to compare this selected model with deep learning models such as the decision tree model, linear discriminant, logistic regression, support vector machine and K-Nearest Neighbor.

Table 10
Comparison of deep learning models

Model	RMSE	MSE	MAE	R^2	Sensitivity	Specificity	Accuracy
GPR-RQ-Zero	0377	0142	0315	0,43	0,83	0,79	0,81
Decision Tree	0539	0290	0290	0,17	0,68	0,73	0,71
Linear Discriminant	0503	0253	0253	0,24	0,78	0,71	0,75
Logistic regression	0506	0256	0256	0,24	0,72	0,77	0,74
Svm	0436	0190	0190	0,38	0,87	0,74	0,81
K-nn	0539	0290	0290	0,17	0,65	0,71	0,68

This outcome was unforeseen, particularly in the context of comparing the GPR model with commonly used deep learning models for classification purposes. It is worth highlighting that the RQ-Zero model demonstrated noteworthy performance, positioning it at the forefront of the rankings alongside the Svm model. This accomplishment is significant, as the RQ-Zero models have demonstrated their capacity to effectively learn and attain appropriate classification outcomes. Additionally, slight distinctions between the Svm and RQ-Zero models have been observed, rendering it challenging to determine the optimal model between them, particularly due to the equivalent classification accuracy they exhibit.

Discussion and conclusion

It was found that it is necessary to carry out more surveys that focus on predicting financial distress using the Gaussian regression method. For this reason, the following text will focus on the results of surveys devoted to predicting the financial crisis of the company through various methods. First, we can immediately mention the study by Jeong and Kim (2022), who designed a model to predict the financial distress of construction companies, considering three, five, and seven years before the prediction point. To construct the prediction model, they chose the financial ratio as an additional input variable, adopted in existing studies of medium- to long-term predictions in other industries. They compared the performance of single-machine and ensemble models to compare the performance of prediction models. A comprehensive comparison of the performance of these models was based on the average value of the prediction performance and the results of the Friedman test. The development of the comparison determined that the random subspace (RS) model showed the best performance in predicting the financial situation of construction companies in the medium to long term.

Rahman *et al.* (2021), in turn, investigated the application of a predictive model of financial distress, which uses the F-score method, including its components, in order to identify companies with a high risk of failure. The dataset was created on the basis of the UCLA-LoPucki Bankruptcy Research database, where 81 publicly traded American companies in financial distress were specifically monitored for the period 2009-2017. The survey concluded that the relationship between the F-score and the likelihood of a company becoming financially distressed is significant. Among other things, the results also show that companies in crisis have negative cash flow from operations (CFO) and show a more significant decrease in return on assets (ROA) in the year before the crash.

As part of their research, the authors Chen and Shen (2020) applied hybrid machine learning methods that integrate stepwise regression, regression and classification trees, selection and the least absolute shrinkage operator, and random forests. They used all these methods in order to create models with which it will be possible to predict the financial distress of the company. The authors used a total of 14 financial variables and 6 non-financial variables for the research. The results show that the CART-LASOO model has the highest level of accuracy, namely 89.74%. We can also mention the study by Chen and Du (2009), who used data mining and neural network clustering to predict financial distress. Here, 33 variables of a financial nature and 4 variables of a non-financial nature were applied to the research. The conclusions of the study show that better accuracy is achieved by models designed using artificial neural networks. In order to predict financial distress, Gregorova *et al.* (2020) method – LR (logistic regression), RF (random forests) and NN (neural networks), using 14 financial ratios. The best performance was assigned to the NN model with an accuracy result of 88.6%. Chen and Jhuang (2020), who also use ANN and CHAID, SR-C5.0 methods, were responsible for another model used to predict financial distress. Using 18 variables of a financial nature and 3 non-financial variables, they found that the SR-C5.0 model showed the highest level of accuracy. The overall accuracy rate was 91.65%. The main goal of Jan's (2021) study was to create highly efficient and accurate models that will be able to predict financial distress using deep neural networks (DNN) and convolutional neural networks (CNN). Based on the results, the authors

Print ISSN: 1390-6291; electronic ISSN: 1390-8618

concluded that the highest financial distress prediction accuracy rate of 94.23% and the lowest type I error rate and type II error rate which are 0.96% and 4.81% respectively.

Thanks to the above results, it is now possible to proceed with the answers to the research questions.

1. Is the GPR model suitable for predicting financial distress?
2. Although, according to the analysis of the existing literature dealing with this issue, the GPR model is not a widely used tool in practice for financial distress, the results of this survey show that the GPR model is excellent for these needs. This is mainly because the model achieves very satisfactory results, with a classification accuracy of 81%.
3. Did the GPR model hold up against the logistic regression model for predicting financial distress?

After comparing the results of this model with deep learning models, respectively, with the linear regression model, it was found that the GPR model outperformed this commonly used model. As mentioned above, the GPR model achieved a classification accuracy of 81%, while the linear regression model achieved only 74%.

In the first phase, we identified the most suitable model among the GPR models by comparing their Kernel functions, and the model was RQ. In the subsequent phase, focused on enhancing the model's performance through hyperparameter optimization, we were able to identify the optimal model from the GPR models based on the variation in Basis function, which was named RQ-Zero. After comparing the results of this model with the results of other deep learning models, we concluded that the model performance was excellent because it achieved very relevant results, as it outperformed all other commonly used models equally with the SVM model, and this prompts us to ask, why has not the GPR model been tried in predicting financial distress based on the difference of the size and type of test sample in a way that makes it commonly used in predicting financial distress or predicting bankruptcy? Moreover, this is even though GPR achieves very relevant results. We also conclude that there is an inverse relationship between the error values and R^2, as the lower the error values, the higher the R2 value. This indicates the accuracy and quality of the model in predicting financial distress, and the opposite is correct. On the other hand, the PCA technique did not achieve the desired objective of its use unusually, as an improvement in the results was achieved after disabling this technique. Finally, we recommend testing the GPR model in predicting financial distress based on a different study sample.

The importance of predicting financial distress using GPR is underscored by the findings of this research paper. GPR has demonstrated a remarkable capacity for accurate prediction, particularly when its hyperparameters are optimized. This model has exhibited superior performance compared to other deep learning models and is on par with Support Vector Machines (SVM), which in itself is a noteworthy achievement. To the best of our knowledge, GPR is an infrequently employed technique, particularly in the context of predicting distress or bankruptcy. Thus, this study aims to alter researchers' perspectives regarding the utilization of GPR in this domain. By exploring novel variations of GPR models and subjecting them to new and diverse study samples, it is possible to identify and address the limitations of previous research, including the present study. Such efforts can expand the outcomes and benefits for all stakeholders involved in this subject, including lenders, auditors, investors, government entities, and particularly companies. This is because the continuity of a company is interconnected with the overall stability of the state's economy. Accurately predicting a company's financial distress facilitates the maintenance of prosperity, minimizes losses, increases investment rates, preserves job opportunities, avoids layoffs, and sustains a mutually beneficial environment for all parties involved.

Support and financial support of research

This research was funded by the Institute of Technology and Business in České Budějovice, the project: IVSUZO2301 - The impact of the circular economy on the share prices of companies listed on the stock exchange.

References

Asante-Okyere, S., Shen, C., Ziggah, Y. Y., Rulegeya, M. M. and Zhu, X. (2018). Investigating the predictive performance of Gaussian process regression in evaluating reservoir porosity and permeability. *Energies*, *11*. https://doi.org/10.3390/en11123261

Bonello, J., Brédart, X. and Vella, V. (2018). Machine learning models for predicting financial distress. *Journal of Research in Economics, 2,* 174-185. https://doi.org/10.24954/JORE.2018.22

Chen, S. and Shen, Z. D. (2020). Financial distress prediction using hybrid machine learning techniques. *Asian Journal of Economics, Business and Accounting*, *16*, 1-12. https://doi.org/10.9734/ajeba/2020/v16i230231

Chen, S. D. and Jhuang, S. (2018). Financial distress prediction using data mining techniques. *ICIC Express Letters, Part B: Applications*, *9*(2), 131-136. https://bit.ly/3qH5eHc

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

Costa, M., Lisboa, I. and Gameiro, A. (2022). Is the financial report quality important in the default prediction? SME Portuguese Construction Sector Evidence. Risks, 10(5). https://doi.org/10.3390/risks10050098

Ferkousl, K., Chellalil, F., Kouzoul, A. and Bekkar, B. (2021). Wavelet-Gaussian process regression model for forecasting daily solar radiation in the Saharan climate. *Clean Energy, 5*(2), 316-328. https://doi.org/10.1093/ce/zkab012

Gavurova, B., Belas, J., Bilan, Y. and Horak, J. (2020). Study of legislative and administrative obstacles to SMEs business in the Czech Republic and Slovakia. *Oeconomia Copernicana, 11*(4), 689-719. https://doi.org/10.24136/OC.2020.028

Gregova, E., Valaskova, K., Adamko, P., Tumpach, M. and Jaros, J. (2020). Predicting financial distress of slovak enterprises: comparison of selected traditional and learning algorithms methods. *Sustainability, 12*(10). https://doi.org/10.3390/su12103954

Hamoudi, Y., Amimeur, H., Aouzellag, D., Abdolraso, M. G. M. and Ustun, T. S. (2023). Hyperparameter bayesian optimization of Gaussian process regression applied in speed-sensorless predictive torque control of an autonomous wind energy conversion system. *Energies*, *16*(12). https://doi.org/10.3390/en16124738

Hantono, H. (2019). Predicting financial distress using Altman score, Grover score, Springate score, Zmijewski score (case study on consumer goods company). *Accountability*, *8*(1), 1-16. https://doi.org/10.32400/ja.23354.8.1.2019.1-16

Herfurth, H. (2020). Gaussian process regression in computational finance. *Project Report, Uppsala University*, 1-29. https://bit.ly/3KGoUSk

Horak, J., Vrbka, J. and Suler, P. (2020). Support vector machine methods and artificial neural networks used for the development of bankruptcy prediction models and their comparison. *Journal of Risk and Financial Management*, *13*(3). https://doi.org/10.3390/jrfm13030060

Jan, C. l. (2021). Financial information asymmetry: using deep learning algorithms to predict financial distress. *Symmetry*, *13*(3). https://doi.org/10.3390/sym13030443

Jeong, J. and Kim, C. (2022). Comparison of machine learning approaches for medium-to-long-term financial distress predictions in the construction industry. *Buildings, 12*(10). https://doi.org/10.3390/buildings12101759

Kliestik, T., Vrbka, J. and Rowland, Z. (2018). Bankruptcy prediction in Visegrad group countries using multiple discriminant analysis. *Equilibrium-Quarterly Journal of Economics and Economic Policy*, *13*(3), 569-593. https://doi.org/10.24136/eq.2018.028

Krulicky, T. and Horak, J. (2021). Business performance and financial health assessment through Artificial Intelligence. *Ekonomicko - manažerské spektrum*, *15*(2), 38-51.

Liew, K. F., Lam, W. S. and Lam, W. H. (2023). Financial distress analysis of technology companies using grover model. *Computer Sciences & Mathematics Forum*, *7*(1). https://doi.org/10.3390/IOCMA2023-14405

Liu, Y., Chen, K., Kumar, A. and Patnaik, P. (2023). Principles of machine learning and its application to thermal barrier coatings. *Coatings, 13*(7). https://doi.org/10.3390/coatings13071140

Paule-Vianez, J. (2019). Bayesian networks to predict financial distress in spanish banking. *Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA, 20*, 131-152. https://doi.org/10.24309/recta.2019.20.2.02

Qu, Y., Quan, P., Lei, M. and Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science, 162,* 895-899. https://doi.org/10.1016/j.procs.2019.12.065

Print ISSN: 1390-6291; electronic ISSN: 1390-8618

Rahman, M., Sa, C. L. and Masud. M. A. K. (2021). Predicting firms' financial distress: an empirical analysis using the F-Score Model. *Journal of Risk and Management, 14*(5). https://doi.org/10.3390/jrfm14050199

Shi, Y. and Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital Journal*, *15*(2), 1866-1875. https://doi.org/10.3926/ic.1354

Taki, M., Rohani, A., Soheili-Fard, F. and Abdeshahi, A. (2018). Assessment of energy consumption and modeling of output energy for wheat production by neural network (MLP and RBF) and Gaussian process regression (GPR) models. *Journal of Cleaner Production, 172*, 3028-3041. https://doi.org/10.1016/j.jclepro.2017.11.107

Vochozka, M., Vrbka, J. and Suler, P. (2020). Bankruptcy or success? The effective prediction of a company's financial development using LSTM. *Sustainability*, *12*(18). https://doi.org/10.3390/su12187529

Wang, S., Gong, J., Gao, H., Liu, W. and Feng, Z. (2023). Gaussian process regression and cooperation search algorithm for forecasting nonstationary runoff time series. *Water*, *15*(11). https://doi.org/10.3390/w15112111

Yang, Z., Li, X., Yao, X., Sun, J. and Shan, T. (2023). Gaussian Process Gaussian Mixture PHD filter for 3D multiple extended target Tracking. *Remote Sensing*, *15*(13). https://doi.org/10.3390/rs15133224

Zhou, T., Song, Z. and Sundmacher, K. (2019). Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. *Engineering, 5,* 1017-1026. https://doi.org/10.1016/j.eng.2019.02.011