Sigma Journal of Engineering and Natural Sciences, Vol. 42, No. 6, pp. XX-XX, December, 2024
Yildiz Technical University Press, Istanbul, Turkey

DOI: 10.14744/sigma.2023.00123

Cryptocurrency price prediction using GPR and SMOTE

Tuğçe GÖKÇEN[1,*], Alper ODABAŞ[2]

ABSTRACT

Cryptography is used by cryptocurrencies to shift money without the intervention of centralized financial institutions. They are decentralized digital assets. On rapidly changing exchanges like those for crypto currencies, it is a tremendously taxing procedure for people to keep track of many simultaneous instantaneous price changes. As a solution to this, computer software that can make fast and objective decisions by constantly observing can replace humans. In this study, the closing price of Bitcoin (BTC), which has the highest volume in the crypto money system, is analyzed. In the study, in which the Gaussian Process Regression (GPR) model and the SMOTE method were used, data belonging to BTC for the period between 25/07/2010 and 05/06/2022 were used as the data set. Opening price, highest-lowest price, volume, dollar index and some indicators used in technical analysis were used as input parameters. The kfold method was followed in the separation of training and test data. The data is divided into 5 subsets with kfold. The mean MAPE value was found to be 1887, and the mean R^2 value was found to be 0.99977 in the models with SMOTE. In addition, the GPR model and the GPR model functions that were applied to the SMOTE method were compared by excluding the opening price, which was the price that was highest-lowest, from the data. It was carried out to determine which model performed better.

Keywords: *Cryptocurrency; GPR; SMOTE; Bitcoin*

INTRODUCTION

Combining the words "crypto" and "currency," "crypto" refers to a cryptocurrency. It is taken from virtual wallets with a password and removed with a password, which is why it is used as a password. Coins such as Bitcoin, Ripple, and Ethereum see the same transaction as real money. One thing that sets it apart from the present coinage is the fact that it is exclusively accessible through the cryptocurrency exchange. Cryptocurrency and blockchain technology are the most important technologies of the digital age. The concept of crypto money, which has entered our lives since 2009, is based on "blockchain" technology. The use of cryptocurrencies has grown exponentially over the years, as more and more people prefer them as an alternative currency. Blockchain technology has allowed the development of many cryptocurrency systems, such as Bitcoin and Ethereum. However, this technology is also applied in many different fields. Cryptocurrencies, which have entered our lives since 2009, have spread all over the world in a short time due to the fact that they are not affiliated with any authority and are anonymous. The digital currency Bitcoin (BTC), the first of the cryptocurrencies, first appeared in 2009 and opened the doors to a new era of crypto money. In order to make anonymous payments completely outside of banks and governments, Bitcoin was created as a so-called virtual money. Bitcoin operates as a middleman in peer-to-peer transactions. It is not to seek assistance from any unit while carrying out this action. As a result, the Bitcoin rate is controlled by the market's supply and demand. Platforms for exchanging virtual money for fiat money emerged with the advent of virtual money. Users that want Bitcoin do so by selling various commodities and services in order to earn Bitcoin. Bitcoin is the most valuable cryptocurrency in the world and is traded on more than 40 exchanges [1]. Unlike other currencies such as the euro or the US dollar, the value of BTC is established by supply and demand, or the price individuals are willing to pay, rather than by a single entity such as a central bank. The price of Bitcoin is governed by supply and demand. When there is a higher demand for Bitcoin, the price rises; when there is a lower demand, the price falls. Request; It is subject to a

This paper was recommended for publication in revised form by Regional Editor Ahmet Selim Dalkilic
[1] Department of Mathematics and Computer Sciences, Faculty of Science, Eskişehir Osmangazi University, Eskişehir, Türkiye
** E-mail address: tugcegokcen95@gmail.com*
Orcid id: https://orcid.org/0000-0003-2655-8363 Tuğçe Gökçen, 0000-0002-4361-3056 Alper Odabaş
Manuscript Received 04 July 2023, Revised 03 August 2023, Accepted 19 September 2023

variety of circumstances, including worldwide events such as price decreases, increases in stock and bond prices, and global economic developments such as the ongoing trade war between the United States and China.

Studies on Bitcoin in the literature have focused on price prediction and cryptology. Using the Grey model, Back Propagation Artificial Neural Network and the Integrated Model of the Grey Neural Network (IMGNN), Zaj et al. [2] attempted to forecast the price of BTC. In comparison to the Grey model and the integrated model, the Backpropagation Artificial Neural Network model has the lowest absolute error rate (5.6%). Shankhdhar et al., [3] aimed to find the least time-consuming and most accurate model for predicting bitcoin price from various machine learning models such as multivariate linear regression, Theil-Sen regression, and Huber regression, as well as deep learning algorithms such as LSTM and GRU. As a result of the research, it was seen that all models had approximately the same accuracy, but the linear regression model had the best execution time. Livieris et al., [4] conducted a comprehensive empirical study using three straight years of crypto data from the three most valuable cryptos, BTC, Etherium (ETH), and Ripple (XRP). For cryptocurrency price and movement prediction, a deep neural network model based on a multi-input architecture is proposed. This same presented forecasting model used these as independent inputs to initially use and process the information from each cryptocurrency separately. The aim of Phaladisailoed & Numnonda, [5] is to find the most efficient and accurate model for predicting Bitcoin prices using various machine learning algorithms. Several distinct regression models were analyzed using trading data collected at 1-minute intervals from the bitcoin exchange website Bitstamp beginning on January 1, 2012, and continuing until January 8, 2018, with the help of the scikit-learn and Keras libraries. The best outcomes had a Mean Squared Error (MSE) of 0.00002 and an R-Squared (R^2) of 99.2%. Madan et al., using machine learning methods and 25 different features related to Bitcoin price, made a price change prediction on five-year Bitcoin price data consisting of daily data. They stated that the performance of the algorithm obtained was 98.7%. They used leverage data at 10-minute and 10-second time points to evaluate Bitcoin price predictions at various levels of precision and noise. It is stated that the performance of the model obtained for the prediction of future price changes in 10-minute time intervals is between 50 and 55% [6].

Technical analysis is the process of forecasting future price movements based on past market movements such as price and volume. Various graphic representations and statistical methods are used for this purpose. It is predicated on the assumption that the formations that appear on the charts as a result of past price movements are indicators for the future, and thus such formations can be used as trading indicators. The aim of the Arslan and Kırcı [7] study in the literature is to generate buy-sell signals using the linear regression method over the highest, lowest, volume, and supply-demand data of bitcoin on the daily chart. In the input data, he used the RSI and MACD indicators. The test's findings revealed that the accuracy percentage was 95.5%. Jiang et al. [8] used logistic regression models to predict the fluctuating trends of stock prices using 19 technical indicators. As a result of the test, the MCP and SCAD accuracy values are 0.732 and 0.731, respectively. MCP and SCAD AUC values were found to be 0.778 and 0.777, respectively. Using an approach that can be referred to as S_I_LSTM and that takes into account a variety of data sources as well as investor sentiment, Wu et al. [9] carried out their stock price estimation. Technical indicators, chronological price information, and non-traditional data sets like as stock notices and economic news.are examples of data sources. Based on the results of the tests, the expected closing price of the stock is closer to the actual closing price, and the mean absolute error could also reach 2.386835. H. Liu's [10] aim was to conduct stock prediction research using the principal component long-short-term memory (LSTM) model. Traditional methods frequently result in poor generalization and predictive effect due to many input data variables, data information overlap, outliers having a large impact on training, and other factors. Given these concerns, it intended to use principal components analysis to reduce the size of the fundamental data before combining the stock-related technical indicators KDJ and MACD as input data and making predictions after adjusting according to the stock characteristic model. The experimental results show that the PCA-S-LSTM model not only reduces mean prediction error but also significantly reduces running time and improves prediction stability.

In their study, Kilimci et al. [11] attempted to predict the final price of the Bitcoin Dollar rate in short-term or frequent trading, known as day trading. A feature set may include technical indicators such as the Bollinger band (BB), hourly moving average (MA), and Relative Strength Index (RSI) in addition to statistical indicators such as maximum, minimum, and average prices (RSI). Convolutional neural networks (CNNs), long short-term memory networks (LSTMs), convolutional long-short-term memory networks (ConvLSTMs), and CNN Long Short-Term

Memory (CNN-LSTM) with various deep and hybrid deep learning methodologies were used to forecast Bitcoin's price. ConvLSTM outperformed the others in predicting bitcoin price, according to the experimental results.

In their study, Sun and Glabadanidis [12] found that technical indicators have significant predictive power on the Chinese stock risk premium. Mohapatra et al. [13] developed aggregated machine learning models to forecast stock returns of Indian banks using some technical indicators. Like Random Forest, XGBoost, Gradient Boosting, and AdaBoost. According to the results, the XGBoost algorithm outperforms the other three ensemble models. It is in the range of 3-5% based on the mean absolute error and the mean root mean square error. Erfanian et al., [14] used machine learning techniques to predict the price of bitcoin (BTC) using both macroeconomic and microeconomic models in their study. They used contrastive techniques such ordinary least squares (OLS), ensemble learning (EL), support vector regression (SVR), and multilayer perceptrons to examine whether macroeconomic, microeconomic, technical, and blockchain indicators based on economic theories are predictive (MLP). The outcomes demonstrate that specific technical indications are crucial for projecting the price of BTC, demonstrating the validity of the technical analysis. Yang et al., [15] used commodities, indicators, and traditional indices to forecast gold prices. They used machine learning algorithms such as Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), and Support Vector Regression (SVR). Two major US indices (the S&P 500 and the DJI), two popular cryptocurrencies (BTC and ETH), and two commodities are included in the dataset (silver and crude oil). The SVR model performed better, and the results revealed that more cryptocurrency data benefited all three models.

When working with data that had strongly skewed class distributions, an issue of how to attain the requisite classification accuracy surfaced in the 1990s as more data and applications of machine learning and data mining became common [16–19]. Class imbalance, known as the unbalanced data problem, is a problem that needs attention within the framework of data science projects. Most classification algorithms are built on the assumption that the training sets are well-balanced. Algorithms are typically designed to maximize the correct prediction rate. However, in real-world datasets, this assumed balanced distribution is not always found. One class can be represented by a small number of instances, whereas the other class can be represented by a large number of instances. Classification issues may arise in this case. In samples with little label information, the model is likely to make erroneous predictions because the model is not adequately trained. Many strategies have been developed to overcome the unbalanced dataset problem. The first method to use when working with an unbalanced dataset is to adjust the class distributions by resampling the data. These methods are undersampling, oversampling, and some sampling techniques. Undersampling aims to rebalance the dataset by eliminating samples of the majority class until the class distributions are equal. The main disadvantage of this method is that it removes from the data set observations that may contribute to projects with an insufficient amount of data points. Also, the randomness of the sample space may suffer if there are few observations. Oversampling increases the number of minority class samples until class distributions are equal. Since most of the methods in this topic copy instances of the minority class, the probability of overfitting increases. Also, in the case of a large dataset with highly uneven distribution, oversampling can be computationally very costly.

Aside from undersampling and oversampling, there are also "advanced sampling techniques" that use exploratory methods to rebalance distributions. In 2002, Chawla, Bowyer, Hall, and Kegelmeyer [20] introduced a novel methodology as a potential substitute for the conventional random oversampling technique. The objective was to address the issue of overfitting by employing oversampling through replication, so aiding the classifier in enhancing its ability to generalize on the testing data. One of them is known as SMOTE, which stands for the Synthetic Minority Over-Sampling Method. It is an oversampling process that makes it possible to produce synthetic data [20]. The method's main goal is to extend the minority class by performing specific actions on existing instances of the minority class. In the field of unbalanced classification, the SMOTE preprocessing technique has become a forerunner. Since its release, numerous extensions and alternatives have been suggested to enhance its efficacy in various scenarios. SMOTE is regarded as one of the most influential data preprocessing/sampling algorithms in machine learning and data mining due to its prominence and impact. In this study, the SMOTE method is used to add more data to the GPR model in order to increase model performance.

The data used in this study consists of the daily price of BTC, one of the cryptocurrencies with the highest market value, from 25/07/2010 to 05/06/2022, obtained from TradingView. The Kfold method was followed in the separation of training and test data. Regression analyses were used as the estimation method. Regression analysis is a

predictive modeling technique that investigates the relationship between dependent and independent variables. The closing price was used as the dependent variable, and the opening price of BTC, the highest-lowest price, volume, dollar index, and some indicator data used in the technical analysis were used as independent variables. As an approach, the Gaussian Process Regression (GPR) model was utilized. These processes were carried out in the MATLAB working environment. In addition, genetic algorithm method was applied in this study. The genetic algorithm (GA) is a mathematical method that operates in a highly parallel manner. It takes a group of individuals, each with a corresponding fitness value, and applies several operators, including reproduction, mutation, and crossover, to generate a new population, known as the next generation [21].

This study compares forecast results using regression analysis and the model functions of the cryptocurrency BTC. The concepts of crypto money, bitcoin, machine learning, regression, and technical analysis are discussed in the first chapter, as are studies of these concepts in the literature. The definition of the variables in the data, the GPR model used, and the function of the SMOTE concept are all explained in the second part. Performance evaluation criteria, analysis results, and graphics are included in the third section. The fourth section contains the test results.

MATERIAL AND METHODS

DATA PREPROCESSING

The data used in this study are the daily USD values of BTC from July 25, 2010, to June 5, 2022, which are available at TradingView. This data set contains a daily worth of data. Dataset characteristics are shown in Table 1.

Table 1. Parameters

Parameter	Explanation
Open	It is the first price of the Bitcoin in the day.
High	It is the highest price of Bitcoin during the day.
Low	It is the lowest price of Bitcoin during the day.
MA	The moving average (MA) is used to calculate the average closing price of an asset over a certain period of time.
Volume	It is the daily trading volume.
RSI	The relative strength index, developed by Welles Wilder Jr. in 1978, is a financial technical analysis indicator. It displays the degree of acceleration of these changes as well as whether a stock is in an uptrend or downtrend over a specific time frame. Although it is not generally used alone, it can give buy-sell signals to the investor over the values it produces. The number of periods in which the RSI is most commonly used is 14. When the analyzed chart is viewed hourly or daily, high closing prices are divided into low closing prices based on the closing values of each hour or day, and the relative strength is calculated. One more of this obtained value is divided by 100 and subtracted from 100, so it always takes a value between 0 and 100. The RSI value is generally evaluated according to the 30-70 bands. Values above 70 indicate that the stock is overbought and should be sold. Values less than 30 indicate oversold conditions and signal a buy signal.
MACD	In the 1960s, Gerald Appel came up with this indicator. It is called the convergence or divergence of the moving averages. It is one of the most used indicators in the technical analysis world.
Bollinger Bands (Basis, Upper, Lower)	Bollinger bands are volatility bands developed by John Bollinger in 1980 and placed above and below the moving averages. They are frequently used in technical analysis. Volatility is a variable dependent on the standard deviation, and increases or decreases in volatility affect the standard deviation. Bollinger bands widen when volatility increases and narrow when volatility decreases. Bollinger bands were patented by John Bollinger in 2011. Bollinger bands show whether prices are relatively high or low. Bands, according to Bollinger [22], account for 88–

	89% of price movements. As a result, he claims that price movements outside of the Bollinger bands are unusual. Technically, prices close to the upper band are considered relatively high, while prices close to the lower band are considered relatively low.
WMA	The weighted moving average (WMA) is a technical analysis tool that can help you determine or confirm the direction of a trend. This indicator is calculated by comparing the current price to prior prices. It gives greater weight to the most recent prices and a lesser weight to previous data points; thus, the importance of the data increases linearly. The main idea of this weighting scheme is that the most recent prices are more reliable for predicting future price fluctuations. WMA is calculated by multiplying each number in the data set by a predefined weight and adding the results.
ROC	The rate of change indicator (ROC) is a momentum oscillator. Calculates the price change between periods as a percentage. ROC takes the current price and compares it to the "n" period's (user-defined) previous price. The calculated value is then plotted and fluctuates above and below a zero line.
Coppock Curve	It is an indication of price momentum over a long period of time that is used to identify large pullbacks and gains in stock market indices. It is calculated as the 10-month weighted moving average of the sum of the index's 14-month and 11-month rates of change.
CCI	The Commodity Channel Index is a momentum-based oscillator that determines when an investment is "overbought" or "oversold."
EFI	Elder's Force Index uses price and volume to calculate the strength of a price movement. The indicator can also be used to identify possible price reversals and corrections. EFI is an oscillator with positive and negative values above and below the zero line.
EMA	The exponential moving average is a tool for tracking price movements in any stock or currency pair.
MFI	The Money Flow Index indicator is a technical analysis tool used to measure trading pressure. This is done by analyzing both price and volume.
DXY	DXY is the symbol and abbreviation for the US dollar index. DXY shows the value against 6 different currencies of 6 different countries: the Euro, Japanese Yen, British Pound, Canadian Dollar, Swiss Franc, and Swedish Krona.
Close	It is the last price of the Bitcoin in the day.

GAUSSIAN PROCESS REGRESSION AND SMOTE

The Gaussian process model is a successful way of machine learning that is more preferred in probabilistic, non-parametric situations to handle nonlinear regression problems [23, 24]. Gaussian processes are named after Carl Friedrich Gauss. Gaussian processes can be seen as an infinite-dimensional generalization of multivariate normal distributions. GPR is a non-parametric model suitable for solving nonlinear regression problems [24]. Gaussian processes are used in statistical modeling, regression to multiple target values, and higher-dimensional mapping analysis [25]. The Gaussian Process Regression technique is utilized to infer and/or predict functions that cannot be calculated analytically. GPR can produce successful results even with small amounts of data and has the ability to quantify uncertainty in predictions. Different covariance functions can be used to determine the most accurate option with GPR [26]. It has four different models with different cores. In its most basic setting, a Gaussian process models a latent function based on a limited set of observations. The Gaussian process can be viewed as the extension of a multivariate Gaussian distribution to an infinite number of dimensions, where any combination of finite dimensions results in a multivariate Gaussian distribution that fully specifies the mean and covariance functions. The choice of the mean and covariance function (also known as the kernel) applies smoothness assumptions to the latent function of interest and determines the correlation between the corresponding observation data points, X as a function of the Euclidean distance, and Y output observations. GP is based on the conversion of prior functions to posterior functions in the Gaussian distribution. The test vectors are estimated using a continuous number of training labels, and the

estimation values at this point can be subject to systematic or random variations. GP defines the probability distribution over functions and can be expressed as given below.

$$f(x) \sim GP\big(m(x), K(x,x')\big) \tag{1}$$

$m(x)$ is the mean, $K(x,x')$ is the covariance function and is expressed as follows:

$$m(x) = E[f(x)] \tag{2}$$

$$K(x,x') = E\left[\big(f(x) - m(x)\big)\big(f(x') - m(x')\big)^T\right] \tag{3}$$

The generally preferred SE in the covariance function is given below:

$$K_{SE}(x,x') = \theta_f^2 exp\left(-\frac{1}{\theta_l^2}\|x - x'\|^2\right) \tag{4}$$

The θ_f and θ_l specified in the covariance function are defined as x-scaling (amplitude) and y-scaling (length scale), respectively.

Both the covariance and the mean make it possible to analyze every combination X of the variables that are being used as input. The covariance matrix is as follows:

$$K := k\big((x_1, \dots, x_n), (x_1, \dots, x_n)\big) \tag{5}$$

$$= \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_{n,}x_1) & \cdots & k(x_n,x_n) \end{bmatrix} \tag{6}$$

The covariance function $k(x,x')$ is usually parameterized with a set of kernel parameters or hyperparameters, θ. $k(x,x')$ is often written as $(k(x,x')|\theta)$ to clearly show the dependence on θ [27]

$$\text{Exponentiel } k\left(x_i, x_j|\theta\right) = \sigma_f^2 exp\left(-\frac{r}{\sigma_1}\right) \tag{7}$$

$$\text{Squared Exponentiel } k\left(x_i, x_j|\theta\right) = \sigma_f^2 exp\left(-\frac{(x_i - x_j)^T(x_i - x_j)}{2\sigma_1}\right) \tag{8}$$

$$\text{Matern 5/2 } k\left(x_i, x_j|\theta\right) = \sigma_f^2\left(1 + \frac{\sqrt{5}r}{\sigma_1^2} + \frac{5r^2}{3\sigma_1^2}\right) exp\left(-\frac{\sqrt{5}r}{\sigma_1}\right) \tag{9}$$

$$\text{Rational Quadratic } k\left(x_i, x_j|\theta\right) = \sigma_f^2\left(1 + \frac{r^2}{2\alpha\sigma_1^2}\right)^{-\alpha} \tag{10}$$

where σ_f is the signal standard deviation, $\sigma_1$is the characteristic length, $r = \sqrt{(x_i - x_j)^T(x_i - x_j)}$, and α is the positive value scale mix parameter. Similarly, there are four basic functions as follows:

$$\text{None } H = [\quad](empty\ matrix) \tag{11}$$

$$\text{Constant } H = I_{n\times 1} \tag{12}$$

$$\text{Linear } H = [1, X] \tag{13}$$

$$\text{Pure Quadratic } H = [1, X, X^2] \tag{14}$$

$$X^2 = \begin{pmatrix} x_{11}^2 & x_{12}^2 & \cdots & x_{1d}^2 \\ x_{21}^2 & x_{22}^2 & \cdots & x_{2d}^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}^2 & x_{n2}^2 & \cdots & x_{nd}^2 \end{pmatrix} \tag{15}$$

The correlation coefficient (R), root mean square error (RMSE), and mean absolute error (MAE) are used to measure how well the proposed GPR models work [28].

SMOTE is an oversampling technique based on the production of synthetic data. It is utilized for classifying datasets with significant class imbalances. The primary objective of SMOTE is to create new minority class instances by combining existing minority class examples [28]. Moreover, SMOTE can conduct regression operations [29]. First, within the context of the SMOTE algorithm, we select a minority class, and then we obtain the K-nearest neighbors, with the default value of K being 5. Every K neighbor is anticipated to be a minority situation. Then, one of these K neighbor samples is picked at random using interpolation. Calculating the difference between the instance of the minority class being considered and the value that the selected neighbor took is the first step in doing the interpolation. This variance is then multiplied by a random number between zero and one and added to the minority group. As a result, a new sample is created [28]. In fact, this process of interpolation chooses a point along the "line segment" between features at random [30].

APPLICATION

The K-fold method was followed in the separation of training and test data. The data is divided into 5 subsets with k-fold. The GPR model was used in the data set (Model1) and compared using the SMOTE method (Model2). In addition, their performances were compared by using kernel functions and basis functions, which are GPR model functions. The performance results of the created models are given in **Error! Reference source not found. Error! Reference source not found.** and **Error! Reference source not found.**. In addition, GPR model (eksGPR called Model3) and SMOTE method and GPR model (eksGPR-SMOTE called Model4) were applied together to the missing data obtained by subtracting the opening prices, highest and lowest prices from the data set, and performance comparison was made. The performance results of the models are given in **Error! Reference source not found.** and **Error! Reference source not found.**.

The calculation results and performances of the created models were evaluated. Correlation analysis (R), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), correlation coefficient squared (R^2), corrected R-squared (Adjusted R^2) and absolute error rates calculation (MAPE), are generally used to determine modeling performance using the formulas provided below. In the model calculation results, MAPE values and AdjRsq, that is, corrected R^2 values, were examined.

$$MAPE = \frac{100}{N}\sum\frac{|e_j|}{|A_j|}$$

$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{n}\sum|e_j|$$

$$AdjRsq = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

$$R = \frac{\sum (xy) - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)2}{n}\right)}}$$

$$R^2 = 1 - \frac{ExplainedVariation}{TotalVariation}$$

$$MSE = \frac{1}{n}\sum e_j{}^2$$

The expression indicated by N and n in the above formulas is the sample size. x and y are samples. A_j is the actual value, e_j is the difference between the actual value and the predicted value, and p is the number of independent variables. According to the regression model results, the MAPE and RMSE values are values with extremely small deviations between the actual and estimated values [31]. MAPE estimation is widely used to evaluate and measure the accuracy of output values by measuring estimated error values. MAPE represents error values as percentages. Forecast models with MAPE values below 10% are called "high accuracy," models with MAPE values between 20% and 50% are called "acceptable," and models with MAPE values above 50% are "false" and "faulty" [32]. AdjRsq adjusted R^2 value is a statistical criterion we use to evaluate the performance of the model in regression analysis. The closer it is to 1, the better the model's performance. The GPR model was created by considering the criteria given in **Error! Reference source not found.** in Model1.

The results obtained from Model1 are given in **Error! Reference source not found.**.

Table 2. Results of GPR Performance

Kernel	Basis Function	R	MAE	MSE	RMSE	R^2	AdjRsq	MAPE
Exponential	Constant	0.96427	1204	1.5111e+07	3883.4	0.92881	0.9273	124.81
Exponential	None	0.96453	1203.8	1.5157e+07	3865.7	0.92918	0.92767	124.94
Exponential	Linear	0.99988	91.557	50655	224.91	0.99976	0.99976	1017.4
Exponential	Pure Quadratic	0.99989	95.885	49086	221.42	0.99977	0.99977	1887
Squared Exponential	Constant	0.8256	468.27	2.5634e+08	9823	-0.24188	-0.26833	670.07
Matern 5/2	Constant	0.99966	96.524	1.4748e+05	355.92	0.99932	0.99931	690.04
Rational Quadratic	Constant	0.87404	398.36	3.9722e+08	9170.4	-0.79373	-0.83194	976.93
Squaredexponential	Optimize Hyperparameters	0.99987	133.12	57713	239.42	0.99973	0.99972	3450.3

The graph of Model1 results is obtained as in **Error! Reference source not found.**.

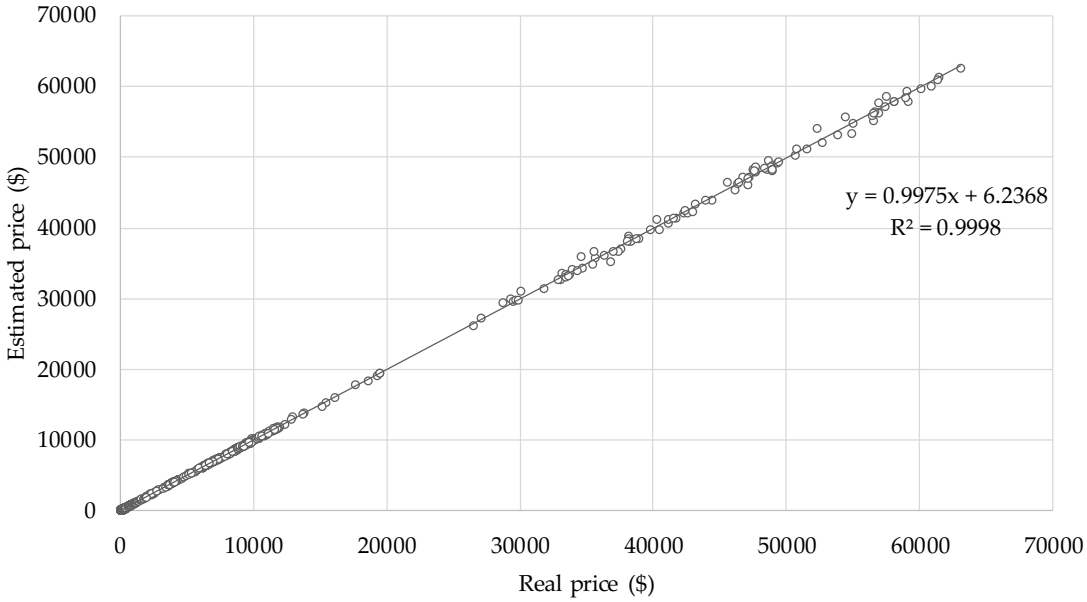

Figure 1. Real versus estimated price for GPR.

Considering the AdjRsq results in **Error! Reference source not found.**, it was seen that the exponential kernel function and pure quadratic basis function model gave better results than the other parameters with a success rate of 0.99977. The correlation graph of the estimated and actual values of the GPR model with the highest performance is as in **Error! Reference source not found.**. The linear graph of the predicted values shown in **Error! Reference source not found.** shows that the performance of the model is high. The y value in the graph shows the equation of the line, and R^2 shows the square of the correlation coefficient. PCA analysis was applied to the data to determine which parameter strongly affected the BTC price and it was seen that the highest parameter was the open price. In addition, optimization was applied to the data. For this, matlab's squaredexponential kernel function and OptimizeHyperparameters are used for the fitrgp model. The AdjRsq result was found to be 0.99972. Genetic algorithm has been applied similarly in various studies in the literature [21, 33–35]. This method could not be applied to Model2, Model3 and Model4.

In order to show the performance of the data obtained in Model2, which is different from Model1, it has been increased synthetically with the SMOTE method. The results obtained from Model2 are given in **Error! Reference source not found.**.

Table 3. Performance of the GPR-SMOTE

Kernel	Basis Function	R	MAE	MSE	RMSE	R^2	AdjRsq	MAPE
Exponential	Constant	0.99037	493.05	3.8697e+06	1965.2	0.98048	0.98027	49.876
Exponential	None	0.99044	490.43	3.845e+06	1958.7	0.9806	0.9804	48.799
Exponential	Linear	0.99991	73.268	34146	183.34	0.99983	0.99983	731.53
Exponential	Pure Quadratic	0.99993	68.307	26951	162.7	0.99986	0.99986	1160.8
Squared Exponential	Constant	0.97825	125.14	1.1231e+07	1791.3	0.94434	0.94375	429.06

Matern 5/2	Constant	0.99984	76.369	63255	236.74	0.99968	0.99968	768.61
Rational Quadratic	Constant	0.97832	125.78	1.1181e+07	1795.2	0.94458	0.944	452.88

The graph of Model2 results is obtained as in **Error! Reference source not found.**.

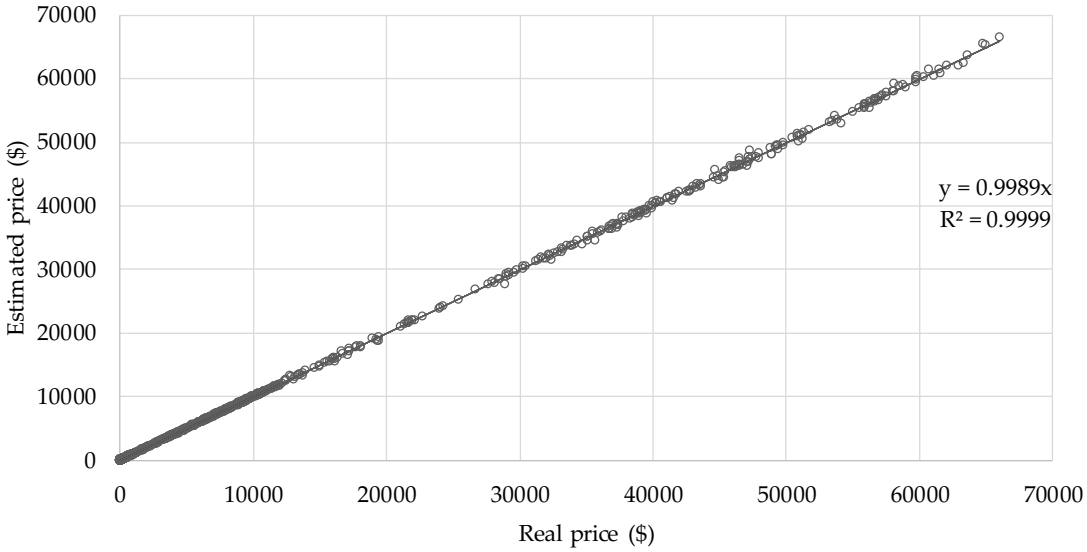

Figure 2. Real versus estimated price for GPR-SMOTE

Considering the AdjRsq results in **Error! Reference source not found.**, it was seen that the exponential kernel function and pure quadratic basis function model gave better results than other parameters, with a success rate of 0.99986. The graph of the GPR model with the highest performance is shown in **Error! Reference source not found.**. The linearity of the graph shows that the model has high performance.

The difference of Model3 from Model1 and Model2, after the opening price, lowest-highest price data were extracted from the obtained data, the GPR model was applied. The results obtained from Model3 are given in **Error! Reference source not found.**.

Table 4. Performance of the eksGPR

Kernel	Basis Function	R	MAE	MSE	RMSE	R^2	AdjRsq	MAPE
Exponential	Constant	0.95582	1389.6	1.8649e+07	4281	0.91301	0.91148	155.16
Exponential	None	0.95854	1374.6	1.7673e+07	4179.5	0.91742	0.91596	155.56
Exponential	Linear	0.99973	140.05	1.1677e+05	340.2	0.99946	0.99945	1329.7
Exponential	Pure Quadratic	0.99976	141.8	1.0264e+05	319.69	0.99952	0.99951	2673.4
Squared Exponential	Constant	0.90436	369.09	5.8318e+07	6367.1	0.72923	0.72944	1085.2
Matern 5/2	Constant	0.96735	205.42	1.7183e+07	2474.7	0.92226	0.92089	931.83
Rational Quadratic	Constant	0.66497	1873.1	4.43e+09	41918	-19.594	-19.958	1243.9

The graph of Model3 results is obtained as in Figure 1.

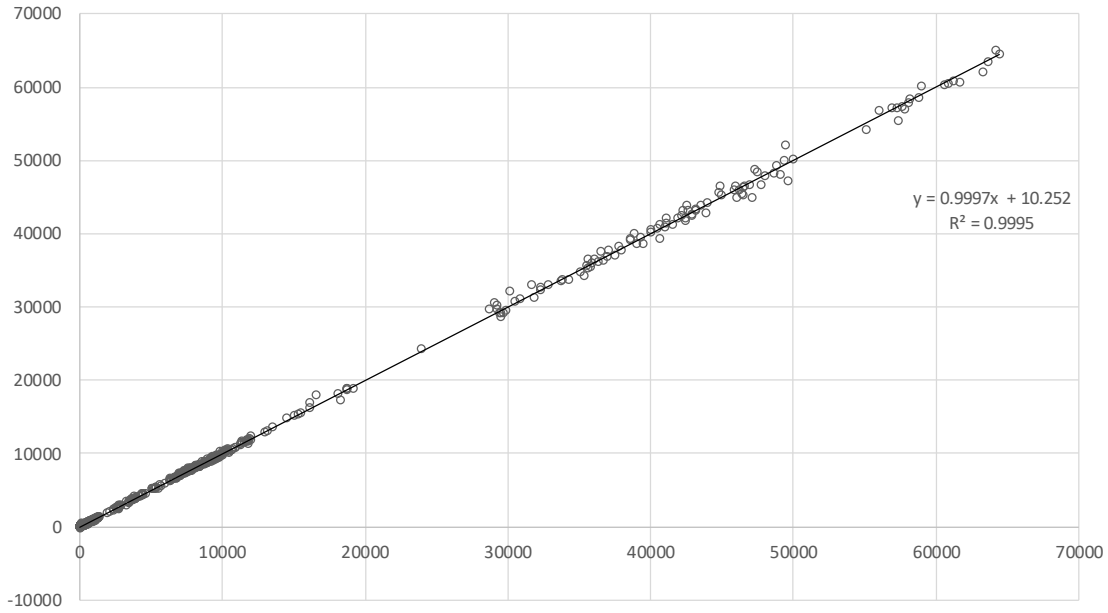

Figure 1. Real versus estimated price for eksGPR

Considering the AdjRsq results in **Error! Reference source not found.**, it was seen that the exponential kernel function and pure quadratic basis function model gave better results than other parameters, with a success rate of 0.99951. The graph of the best performing GPR model is shown in Figure 1. Since the correlation graph is a linear graph, it is seen on the graph that the model performance is good.

Model4 was obtained by creating synthetic data with the SMOTE method in the data of Model3. **Error! Reference source not found.** displays the results obtained from Model4.

Table 5. Performance of the eksGPR-SMOTE

Kernel	Basis Function	R	MAE	MSE	RMSE	R2	AdjRsq	MAPE
Exponential	Constant	0.98857	567.21	4.5819e+06	2137.7	0.97683	0.97662	61.1
Exponential	None	0.98864	565.85	4.5616e+06	2132.7	0.97693	0.97672	60.729
Exponential	Linear	0.99989	82.871	45725	213.51	0.99977	0.99977	1081.7
Exponential	Pure Quadratic	0.99988	88.144	47998	217.42	0.99976	0.99976	1902.7
Squared Exponential	Constant	0.91486	201.34	9.4034e+07	4516.9	0.53018	0.52606	913.84
Matern 5/2	Constant	0.9778	141.54	1.1194e+07	1683.5	0.94407	0.94358	1177
Rational Quadratic	Constant	0.91562	206.95	9.1081e+07	4591.7	0.54489	0.5409	1033.9

The graph of Model4 results is obtained as in Figure 2.

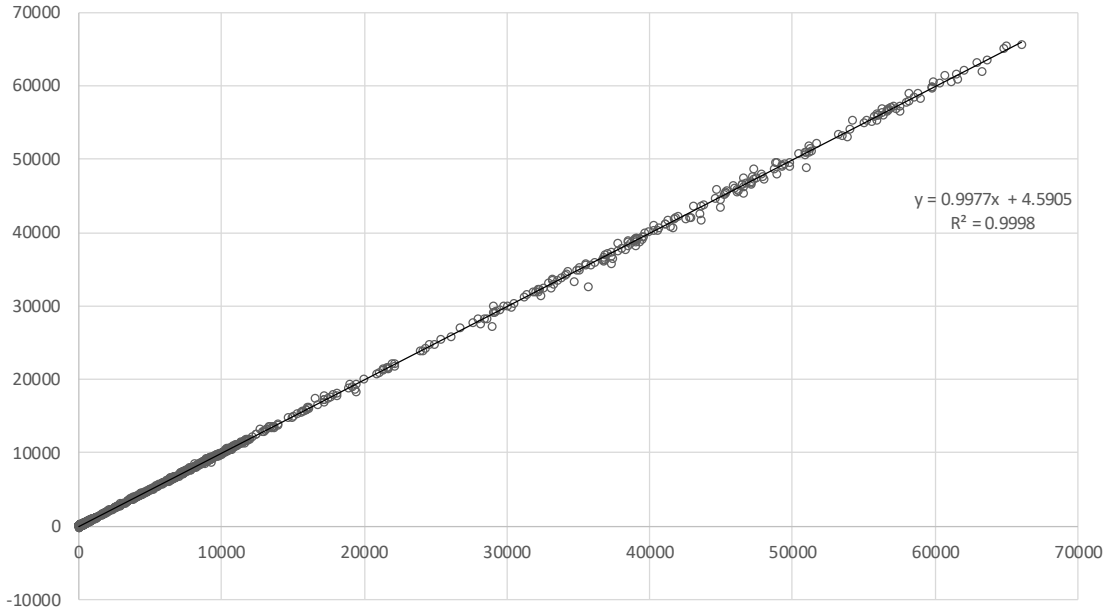

Figure 2. Real versus estimated price for eksGPR-SMOTE

Considering the AdjRsq results in **Error! Reference source not found.** , it was seen that the exponential kernel function and linear basis function model gave better results than other parameters with a success rate of 0.99977. The graph of the best-performing GPR model is shown in Figure 2.

When we look at the results in Model1, Model2, Model3 and Model4, exponential kernel function and pure quadratic basis function model gave the best performance in Model1, Model2 and Model3, and the exponential kernel function and linear basis model gave the best performance in Model4. Although the linear function gives the best performance in Model4, it is seen that it has close performance in the pure quadratic function in **Error! Reference source not found.**. In general, in all models, it was seen that the model gave better results with the exponential kernel function, the pure quadratic basis function, and the synthetically SMOTE method. It is concluded that more data gives better performance.

CONCLUSİON

In our study, a review was made of BTC, one of the cryptocurrencies. The concepts related to the scattered structures and indicators underlying the development of cryptocurrencies are discussed.

The goal of the study is to make a model that can use past data to accurately predict how much cryptocurrencies will cost in the future. In this context, the GPR model was used by applying the Gradient Boosted Trees (GPR) model and SMOTE method for forecasting using the prices of the BTC cryptocurrency between July 25, 2010, and June 5, 2022. It was trained and tested by dividing it into five subsets using Kfold. When the performance criteria MAPE, RMSE, MAE, AdjRsq, R, R^2, and MSE were examined, it was observed that the model using SMOTE was successful on the existing data.

REFERENCES

[1] McNally S, Roche J, Caton S (2018) Predicting the Price of Bitcoin Using Machine Learning. Proc - 26th Euromicro Int Conf Parallel, Distrib Network-Based Process PDP 2018 339–343. https://doi.org/10.1109/PDP2018.2018.00060

[2] Zaj MM, Samavi ME, Koosha E (2022) Measurement of Bitcoin Daily and Monthly Price Prediction Er-ror Using Grey Model, Back Propagation Artificial Neural Net-work and Integrated model of Grey Neural Network. Adv Math Fin App 2022:535–553. https://doi.org/10.22034/AMFA.2020.1881110.1315

[3] Shankhdhar A, Singh AK, Naugraiya S, Saini PK (2021) Bitcoin Price Alert and Prediction System using various Models. IOP Conf Ser Mater Sci Eng 1131:012009. https://doi.org/10.1088/1757-899x/1131/1/012009

[4] Livieris IE, Kiriakidou N, Stavroyiannis S, Pintelas P (2021) An Advanced CNN-LSTM Model for Cryptocurrency Forecasting. https://doi.org/10.3390/electronics

[5] Phaladisailoed T, Numnonda T (2018) Machine learning models comparison for bitcoin price prediction. In: Proceedings of 2018 10th International Conference on Information Technology and Electrical Engineering: Smart Technology for Better Society, ICITEE 2018. Institute of Electrical and Electronics Engineers Inc., pp 506–511

[6] Madan I, Saluja S, Zhao A, et al (2015) Automated Bitcoin trading via machine learning algorithms. Available via DIALOG. Weizmann Inst Sci 1–5

[7] Arslan ME, Kırcı P (2021) Makine Öğrenmesi İle Borsa Analizi. Eur J Sci Technol 1117–1120. https://doi.org/10.31590/ejosat.1012785

[8] Jiang H, Hu X, Jia H (2022) Penalized logistic regressions with technical indicators predict up and down trends. Soft Comput. https://doi.org/10.1007/s00500-022-07404-1

[9] Wu S, Liu Y, Zou Z, Weng TH (2022) S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. Conn Sci 34:44–62. https://doi.org/10.1080/09540091.2021.1940101

[10] Liu H (2021) A research on stock forecasting based on principal component LSTM model. 2021 IEEE Int Conf Adv Electr Eng Comput Appl AEECA 2021 684–688. https://doi.org/10.1109/AEECA52519.2021.9574128

[11] Kilimci H, Yildirim M, Kilimci ZH (2021) The Prediction of Short-Term Bitcoin Dollar Rate (BTC/USDT) using Deep and Hybrid Deep Learning Techniques. ISMSIT 2021 - 5th Int Symp Multidiscip Stud Innov Technol Proc 633–637. https://doi.org/10.1109/ISMSIT52890.2021.9604741

[12] Sun M, Glabadanidis P (2022) Can technical indicators predict the Chinese equity risk premium? Int Rev Financ 22:114–142. https://doi.org/10.1111/irfi.12344

[13] Mohapatra S, Mukherjee R, Roy A, et al (2022) Can Ensemble Machine Learning Methods Predict Stock Returns for Indian Banks Using Technical Indicators? J Risk Financ Manag 15:. https://doi.org/10.3390/jrfm15080350

[14] Erfanian S, Zhou Y, Razzaq A, et al (2022) Predicting Bitcoin (BTC) Price in the Context of Economic Theories: A Machine Learning Approach. Entropy 24:1–29. https://doi.org/10.3390/e24101487

[15] Yang J, De Montigny D, Treleaven P (2022) ANN, LSTM, and SVR for Gold Price Forecasting. 2022 IEEE Symp Comput Intell Financ Eng Econ CIFEr 2022 - Proc. https://doi.org/10.1109/CIFEr52523.2022.9776141

[16] Provost F (2000) Machine Learning from Imbalanced Data Sets 101 Extended Abstract

[17] Zewdu T (1998) Prediction of HIV Status in Addis Ababa using Data Mining Technology

[18] López V, Fernández A, García S, et al (2013) An insight into classification with imbalanced data : Empirical results and current trends on using data intrinsic characteristics. Inf Sci (Ny) 250:113–141. https://doi.org/10.1016/j.ins.2013.07.007

[19] He H, Garcia EA (2009) Learning from Imbalanced Data. 21:1263–1284

[20] Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE : Synthetic Minority Over-sampling Technique. 16:321–357

[21] Arqub OA, Abo-Hammour Z (2014) Numerical solution of systems of second-order boundary value problems using continuous genetic algorithm. Inf Sci (Ny) 279:396–415. https://doi.org/10.1016/j.ins.2014.03.128

[22] Bollinger J (1992) Using Bollinger Bands. Stock Comodities 10:47–51

[23] Aci M, Dogansoy GA (2022) Demand forecasting for e-retail sector using machine learning and deep learning methods. J Fac Eng Archit Gazi Univ 37:1325–1339. https://doi.org/10.17341/gazimmfd.944081

[24] Liu K, Hu X, Wei Z, et al (2019) Modified Gaussian Process Regression Models for Cyclic Capacity Prediction of Lithium-Ion Batteries. IEEE Trans Transp Electrif 5:1225–1236. https://doi.org/10.1109/TTE.2019.2944802

[25] Zhang N, Xiong J, Zhong J, Leatham K (2018) Gaussian process regression method for classification for high-dimensional data with limited samples. 8th Int Conf Inf Sci Technol ICIST 2018 358–363. https://doi.org/10.1109/ICIST.2018.8426077

[26] Heo Y, Zavala VM (2012) Gaussian process modeling for measurement and verification of building energy savings. Energy Build 53:7–18. https://doi.org/10.1016/j.enbuild.2012.06.024

[27] Eberhard J, Geissbuhler V (2000) Konservative und operative therapie bei harninkontinenz, deszensus und urogenital-beschwerden

[28] Pat S (2022) Optik Optical properties of Nb 2 O 5 doped ZnO nanocomposite thin film deposited by thermionic vacuum arc. 258:. https://doi.org/10.1016/j.ijleo.2022.168928

[29] Zhang Y, Xu X (2020) Yttrium barium copper oxide superconducting transition temperature modeling through gaussian process regression. Comput Mater Sci 179:109583. https://doi.org/10.1016/j.commatsci.2020.109583

[30] So B, Boucher J (2021) Synthetic Dataset Generation of Driver Telematics. 1–19

[31] Chen KY, Wang CH (2007) Support vector regression with genetic algorithms in forecasting tourism demand. Tour Manag 28:215–226. https://doi.org/10.1016/j.tourman.2005.12.018

[32] Metin S (2021) Kripto Para Fiyatlarının Regresyon Analizi Yöntemleri ile Tahmini: Bitcoin, Etherum ve Ripple. In: 2. Uluslararası Sosyal Bilimler ve İnovasyon Kongresi

[33] Abo-Hammour Z, Alsmadi O, Momani S, Abu Arqub O (2013) A genetic algorithm approach for prediction of linear dynamical systems. Math Probl Eng 2013:. https://doi.org/10.1155/2013/831657

[34] Abu Arqub O, Abo-Hammour Z, Momani S, Shawagfeh N (2012) Solving singular two-point boundary value problems using continuous genetic algorithm. Abstr Appl Anal 2012:. https://doi.org/10.1155/2012/205391

[35] Abo-Hammour Z, Abu Arqub O, Momani S, Shawagfeh N (2014) Optimization solution of Troesch's and Bratu's problems of ordinary type using novel continuous genetic algorithm. Discret Dyn Nat Soc 2014:. https://doi.org/10.1155/2014/401696