International Journal of Computational Economics and Econometrics

ISSN online: 1757-1189 - ISSN print: 1757-1170
https://www.inderscience.com/ijcee

Machine learning-based business risk analysis for big data: a case study of Pakistan

Mohsin Nazir, Zunaira Butt, Aneeqa Sabah, Azeema Yaseen, Anca Jurcut

DOI: 10.1504/IJCEE.2023.10057679

Article History:

Received:	16 December 2020
Last revised:	26 April 2021
Accepted:	02 September 2021
Published online:	20 December 2023

Machine learning-based business risk analysis for big data: a case study of Pakistan

Mohsin Nazir, Zunaira Butt* and Aneeqa Sabah

Lahore College for Women University,
Jail Road, Lahore, Pakistan
Email: mohsinsage@gmail.com
Email: zunirabutt@gmail.com
Email: aneeqas29@gmail.com
*Corresponding author

Azeema Yaseen

Maynooth University Maynooth,
Co. Kildare, Ireland
Email: azeema.yaseen.2020@mumail.ie

Anca Jurcut

University College Dublin,
Belfield, Dublin 4, Ireland
Email: anca.jurcut@ucd.ie

Abstract: In finance, machine learning helps the business by improving its abilities and flexibility to prevent risks, errors and to accept such challenges. This research analyses and forecasts the interest rate risk of Pakistan using machine learning models. It took a ten-year financial dataset of Pakistan investment bonds from the State Bank of Pakistan website. In this study, a framework was proposed and four different models were developed to forecast the interest rates: neural network, bootstrap aggregated regression trees, cascade-forward neural network, and radial basis neural network. Subsequently, these models were run under four different scenarios: forecasting with original, generated, LASSO extracted and weighted average features. In addition, the outcomes of these models were compared with four performance metrics: mean absolute percentage error, daily peak mean absolute percentage error, mean absolute error, and root mean square error. Overall, the results showed that radial basis neural network provided the best forecasting.

Keywords: machine learning; business risk analysis; interest rate risk; risk analysis; big data; forecasting models; Pakistan.

Reference to this paper should be made as follows: Nazir, M., Butt, Z., Sabah, A., Yaseen, A. and Jurcut, A. (2024) 'Machine learning-based business risk analysis for big data: a case study of Pakistan', *Int. J. Computational Economics and Econometrics*, Vol. 14, No. 1, pp.23–41.

Biographical notes: Mohsin Nazir is the Department Head of Software Engineering at the Lahore College for Women University, Pakistan, and an adjunct faculty member. He holds a PhD and MS from the Asian Institute of Technology, Thailand, and a BS from the National University of Computer and Emerging Sciences, Pakistan. With extensive experience since 2003, his research focuses on the application of information and communications technologies for real-time applications. He has numerous publications, serves on editorial boards, and has organised/participated in national and international conferences. Additionally, he worked as a Research Scientist at the University of Oulu, Finland.

Zunaira Butt is a Computer Scientist pursuing a PhD in Computer Science at the Lahore College for Women University (LCWU), Pakistan. With a Bachelor's in Computer Science in 2013 and a Master's in the same field in 2018 from LCWU, she currently serves as a Lecturer in the Department of Computer Science. She specialises in artificial intelligence, machine learning, deep learning, IoT, and robotics. Her research aims to advance these areas and develop innovative solutions with transformative potential. With dedication and passion, Zunaira contributes significantly to the academic community at LCWU and shows promise as a researcher in computer science.

Aneeqa Sabah is an Assistant Professor in the Physics Department at LCWU. She received an overseas scholarship from HEC for her successful PhD project in nanotechnology. She has supervised numerous projects at the BS, MS, and PhD levels, focusing on nano fabrication, quantum dots, spectroscopy, and more. Her research papers have been published in high-impact journals such as the *Journal of Physical Chemistry* and *ACS*. She also actively participates in seminars, conferences, and organisational activities. Her expertise lies in nano synthesis, self-assembly, colloids, one-dimensional nano materials, green chemistry, and quantum dots.

Azeema Yaseen is a third-year PhD candidate in Computer Science at the Maynooth University, Ireland. She holds a Bachelor's in 2015 and Master's in 2017 in Computer Science from the Lahore College for Women University (LCWU), Pakistan. Her research focuses on the applications of the internet of musical things (IoMusT) and human-computer interaction, particularly in developing interface and interaction frameworks for amateur users in musical contexts. Her work also extends to multimodal interactions and digital wellbeing systems. She has previous experience as a Lecturer at the University of Gujrat (UoG) and the University of Management and Technology (UMT), Lahore.

Anca Jurcut is an Assistant Professor at the School of Computer Science, University College Dublin (UCD), Ireland, since 2015. She holds a BSc in Computer Science and Mathematics from the West University of Timisoara, Romania in 2007, and a PhD in Security Engineering from the University of Limerick (UL), Ireland in 2013, funded by the Irish Research Council for Science Engineering and Technology. With expertise in security protocols, formal verification, network security, IoT security, and blockchain applications, she has made significant contributions to attack detection, prevention techniques, security protocols, and mobile edge computing (MEC).

1 Introduction

Risk management is a vast and vital topic that is being researched in the past few years. In general, risk means a threat or an uncontrollable situation that results in loss. Business risk is an unexpected change in different factors like government regulations, customer demand, customer satisfaction, and competition, etc. that lowers its profit or leads to a major loss. There are many reasons for the risk and delay in the identification of these reasons may encounter big problems. Sometimes risk arises due to taking wrong decisions or unexpected changes in the rules and regulations (Kim, 2019). In the economic world, interest rates (IR) play an imperative role and an integral part of the banking business. In many industries, it is the source of profit. When this risk reaches abnormal levels, it raises the threat of business destruction. Therefore, the stability of interest rate risk is the top priority of any business need. Interest rate risk is the possibility of decreasing rate of an asset resulting from unpredicted and sudden deviations in interest rates (Kumar, 2014).

- assets value reduces
- business profit decreases
- sales value falls
- loans become expensive.

Pakistan is a democratic country in which after every 5 years the government changes. The new government introduces the new policies and framework to improve the economic development of the country, e.g., taxation policies, interest rate policies, regulations, and permits. These policies influence the business directly or indirectly that is very challenging for the business persons (Kim, 2019). This research concentrates on the Pakistan interest rate risk using machine learning (ML) models. It presents a framework and four ML models to analyse, predict and forecast the future of business in terms of interest rates. There are seven different types of interest rate risk whereas this paper focused on one of the interest rate risk type namely yield curve risk. Following are the objectives of the study that explains:

- Why is the interest rate important to study?
- How it affects business growth?
- How machine learning will resolve these issues?
- Which machine-learning models will be best to analyse, predict, and forecast the interest rates?
- How forecasting results will improve the business scope?

In the last few years, ML has been used in many applications such as health, banking, agriculture, economics, weather, big data, etc. and provide very encouraging results for the development of better forecasting system (Amjady and Keynia, 2009; Boyacioglu et al., 2009; Carbonneau et al., 2008; Dibike and Solomatine, 2001; Harrou et al., 2019; Ince and Aktan, 2009; Mehmood and El-Hawary, 2014; Sun et al., 2008; Tavana et al., 2018; Voyant et al., 2017, 2018; Zhang, 2017). Like other applications ML

also integrates in the business sector to forecast interest rate risk. Several studies for instance (Enke and Mehdiyev, 2013; Joseph et al., 2011; Kanevski et al., 2008; Kumar, 2014; Nunes et al., 2019; Oh and Han, 2000; Sambasivan and Das, 2017) adopt machine learning models for IR forecasting. Similarly, in different business areas such as investment, insurance or retail banking big data also helps in risk management. Such as in the investment sector, it can help in the selection of good investments against profit and losses. Likewise, big data also supports by organising data in structured form to recognise risks, their impact ratio, and mitigation rate in an organisation. In short, big data supports organisational decision-makers by identifying different types of risks and their influence on business.

Firstly this paper observed the literature on yield curve forecasting like, Oh and Han (2000) configured a three stages artificial neural network (ANN) model using change point detection technique for interest rate forecasting. At the first stage, the model detected the change in interest rate values. Secondly, backpropagation neural network (BPNN) forecast the change point values and in the last step BPNN showed the forecasting results of interest rates. However, the provided solution only covers short term IR data and may not be conventional for long term IR. So it still needs more research to figure out this point. To forecast yield curve risk, Kanevski et al. (2008) implemented geostatistical models and machine learning models [multilayered perceptron and support vector machine (SVM)] on the Swiss franc interest rate dataset. The full explanation of the geostatistical model is available at Wackernagel (2003). Joseph et al. (2011) used an ANN model to forecast the future results of the 10 years US Treasury bonds and 3-month US Treasury bills. Likewise, Enke and Mehdiyev (2013) suggested type-2 fuzzy clustering and type-2 fuzzy inference neural network model for short term interest rate forecasting. Sambasivan and Das (2017) proposed a Gaussian process framework for yield curve forecasting and its performance compared with the multivariate time series method. In the end, it was observed that the multivariate time series method worked well in the short-term yield curve dataset and the Gaussian process framework worked well in medium and long-term yield curve datasets. The above papers have tended to focus on IR forecasting and the area of big data remains unclear.

Secondly this document review the proposed work on big data like, Bharill et al. (2016) used Apache Spark as a platform and presented a scalable random sampling with iterative optimisation fuzzy c-means (SRSIO-FCM) model that created suitable clusters for big datasets with less computational cost. It was compared with literal fuzzy c-means (LFCM) and random sampling plus extension fuzzy c-means (rseFCM) and results found that SRSIO-FCM took less time in performing efficient clustering. Sakr et al. (2018) described the complete road map and explained the different tools and techniques of big data for business. This study recommended a business process model and developed a framework for big data. Furthermore, it also explained different algorithms and platforms to improve the business model. Ramírez-Gallego et al. (2017) organise a research on six large-scale datasets based on real-life massive data streams. They implemented the nearest neighbour (NN) algorithm on Apache Spark platform to speed up the search. The researchers also suggested an additional selection method for these datasets which continuously updated and deleted the old data from the datasets. Authors claim that it was the first research that provided an efficient solution for high speed streaming big data.

Suleiman et al. (2017) conducted an experimental study using two platforms H2O and Sparkling water on Santander Bank dataset. This research aim was to provide an empirical evaluation of these two emerging platforms and results showed that H2O platform performed best. Tripathy et al. (2017) introduced a new version of support vector machine (SVM) for the risk analysis which is called parallel support vector machine (PSVM). PSVM solved the big data handling issue which occurred in standard SVM. Assefi et al. (2017) conducted research on big data using Apache Spark MLlib platform. Authors used six large datasets to measure the performance of the tool and compared its results with the Weka tool. At the end, it is found that Apache Spark MLlib is a very strong tool for big data than Weka because Weka worked slower under big and heavy datasets. The Following researches (Assefi et al., 2017; Bharill et al., 2016; Ramírez-Gallego et al., 2017; Sakr et al., 2018; Suleiman et al., 2017; Tripathy et al., 2017) delivered the idea of big data handling techniques and ease the implementation of this work.

The core of the study is partially based on Nunes et al. (2019) in which five different models were integrated on European Government bonds dataset. Multivariate linear regression (MLR) and multi-layer perceptron (MLP) were two main models and these two models were used to design the remaining three. However, the results showed that MLP extracted the most relevant features and helped to improve prediction accuracy. The literature review on big data and yield curve risk provided helpful input to conduct this research. In the previous studies, no competent solution is available for yield curve forecasting using machine learning and big data techniques. Here the need of a framework arises that can cover both big data handling and risk analysis problems to forecast appropriate results. This paper proposed a framework consisting of eight stages, including all general steps for evaluating a good forecaster. Furthermore, it can handle any financial data using both big data and machine learning techniques. From the literature review, artificial neural networks stand out as forecasting tools due to its flexibility for a wide number of features. Following ML models were reported best results in diverse markets: ANN, SVM, BPNN, and MLP. In addition, this research aim was to dig out more ML models for yield curve risk forecasting. Therefore, it evaluates artificial neural network (ANN), cascade forward neural network (CFNN), radial basis neural network (RBNN) and bootstrap aggregated regression trees (BART) as forecasters. ANN model was adopted from literature while others were first used for yield curve forecasting.

2 Research method

The focus of this work is the government bond assets class, which has the following reasons. First, its dataset provides information about Pakistan's economic market. Second, it tells about the value of Pakistan bonds in the international marketplace. Further, this study used multivariate time series method to get good forecasting results. The targets to be predicted were: 3, 5, 10, and 20 years. Extracting meaningful features from a large dataset is a very important step to increase the performance of predictive models. The value of a business in markets is based on the number of features on which they measure and compare with each other. Economic variables are very important for the establishment of appropriate yield macro models. The list of original features and their variables are shown in Table 1.

Table 1 List of original features

Feature name	*Feature variables*	
Daily average	3, 5, 10 days	
Monthly average	3, 5, 10 months	
Yearly average	Cut-off yield	3, 5, 7, 10, 15, 20, 30 years
	Weighted average yield	3, 5, 7, 10, 15, 20, 30 years

Table 2 List of generated features

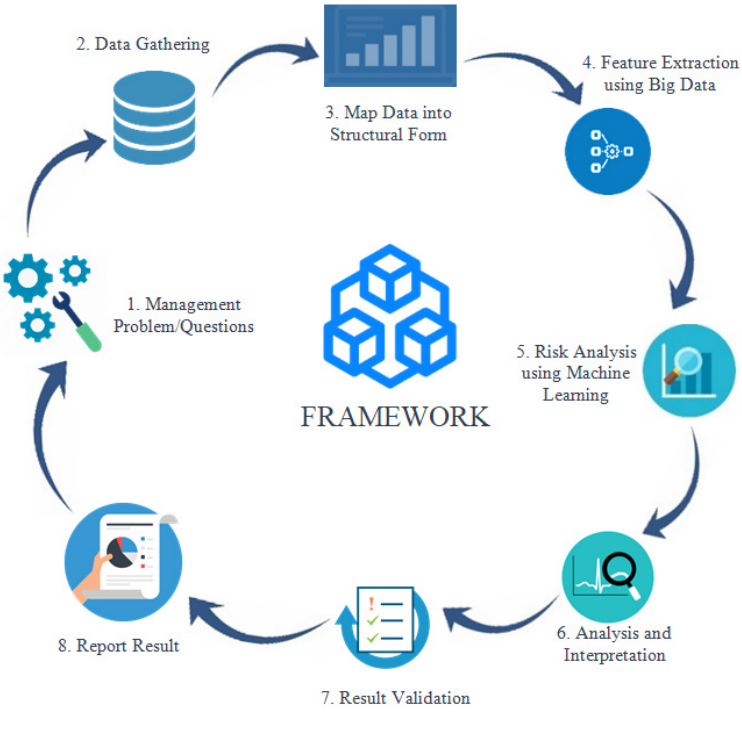
Feature name	*Feature variables*
Spread curve analysis	3M[a]10Y[b], 3Y10Y, 10Y30Y
Technical analysis	3Y10Y, 5Y10Y, 10Y30Y

Notes: [a]M indicates months.
[b]Y indicates years.

Table 3 Summary of features

Original features	20
Generated features	6
Target features	4
Total features	30

Figure 1 Framework for interest rate risk forecasting (see online version for colours)



There is a natural temporal order in financial time series data. This order cannot be changed or disrupted during modelling because the order is itself information in time series. Data period in series and correlation between them is also the major factor of time series which cannot be ignored. So, the new features generated by using original features and past values of time series. Table 2 illustrates the generated features and Table 3 displays the whole summary of features used in this paper. Figure 1 demonstrates the proposed framework for interest rate risk forecasting.

2.1 Management problem/question

First phase of the framework illustrates the number of questions related to the research. At this point multiple features and parameters are set to complete the research goal. These questions are the following:

- Why is the forecast model necessary?
- How results will improve business performance?
- Which type of interest rates data are going to use in the model?
- Is this model handling both long-term and short-term interest rates data?
- Which types of methods are going to adopt for result validation?
- What is the accuracy rate of the result?
- How will the results be validated?

2.2 Data gathering

The dataset was obtained from the State Bank of Pakistan website, which covered the data of Pakistan investment bonds from May 2006 to November 2016 (SBP, 1956). The interval of predicted and forecasted data was from 2017 to 2020. This dataset also contained different variables as parameters and different grouping of bids (bids received, rejected and not received).

2.3 Map data into structural form

When data is gathered from different sources it is very messy and in an unstructured format. Therefore, it converts into a structural form to analyse, explore, and extract important information from the data. Most organisations use a very famous tool 'MS Excel' to organise data. In this research, this tool is used for structuring the dataset. After data gathering, preprocessing applied to convert the unbalanced data into a useful and meaningful format. Often the original data is not in the desired shape and uses this data without preprocessing can raise the issues of out of range values, missing values, incompetent classification, and impossible data combinations, etc. Preprocessing helps the model to generate more good and accurate results. This study applied preprocessing on big data to avoid these issues. The collected data contained redundant values, which raised the redundancy issues and gave inaccurate results. Therefore, these values were removed to clean the disordered data. Additionally, there were missing values in data

that generated the problem in results, so it was essential to handle these missing values and recover them to improve the result performance. For this purpose, the linear interpolation method was used for missing data handling (Abdullah, 2014).

2.3.1 Linear interpolation

This is the simplest method to interpolate the missing values in which two data points are connected with a straight line. It takes more memory and computational time than other interpolation methods. Linear interpolation can be computed by the following equation (Noor et al., 2008).

$$f_1(x) = b_0 + b_1(x - x_0) \tag{1}$$

where x and x_0 are independent variables, x_0 is known value, $f_1(x)$ is the dependent variable for value x. From equation (1).

$$b_0 = f(x_0) \tag{2}$$

and

$$b_1 = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)} \tag{3}$$

2.4 Feature extraction

The purpose of big data is the extraction of meaningful information from the big and massive datasets. When data is converted into the structural form, it shows multiple variables and parameters in the data. To improve the result performance it's essential to find out the relevant and most useful features from data. At this stage, different big data tools and algorithms are used to achieve the goal (Nunes et al., 2019).

2.4.1 LASSO regularisation technique

Least absolute shrinkage and selection operator (LASSO) is the most popular machine learning technique for feature extraction. It is a regression analysis method that performs both regularisation and variable selection tasks to increase the prediction accuracy and interoperability of the model. This study used LASSO using a 10 fold cross-validation technique to extract the useful and most relevant features from big data. Cross-validation is a technique that divides the large dataset into small equal parts and randomly splits them for the training and testing process. It regularises model parameters by eliminating irrelevant features (set feature value to zero) from the dataset. Then this algorithm uses L1 regularisation technique for automatically feature selection and splits useful features based on the criteria having minimum mean square error (MSE). LASSO uses the following equation to regularise the problem (Nunes et al., 2019).

$$\min_{\beta_0,\beta} \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{4}$$

Here,

- N is the number of observations
- y_i is the response at observation i
- x_i is data, a vector of p values at observation i
- λ is a non-negative regularisation parameter corresponding to one value of lambda
- the parameters β_0 and β are a scalar and a vector of length p, respectively
- as λ increases, the number of non-zero components of β decreases.

2.4.2 Cross validation

Cross validation is a model used to overcome the overfit and underfit risk and improves the machine learning models prediction results. K-fold usually 10 folds is the most common technique for cross validation which splits the original dataset into training and testing sets. The training set is used to train the model while testing set is to analyse its performance. Also, this process repeats multiple times and the average of cross validation is used to measure the performance (Nunes et al., 2019).

2.5 Risk analysis using machine learning

After the extraction of useful features the next stage was risk analysis. Four different scenarios were used to forecast interest rates: Forecasting with original, generated, LASSO and weighted average features. Following tools and algorithms were adopted to analyse and forecast risk.

2.5.1 MATLAB

It is the most widely used tool that provides many toolboxes to handle risk management problems. It provides the facility of built-in models and custom models for risk analysis. Also, a financial toolbox is available for mathematical modelling and statistical analysis which offers solutions of risk estimation, analysing the level of interest rate, measuring investment performance, and interest rate derivatives (Brandimarte, 2006).

2.5.2 Artificial neural network

The concept of ANN is inspired by biological neurons in which artificial neurons are connected with each other and creates a whole network. There are three main layers of ANN: input, output, and hidden layers. The number of hidden layers can be increased according to the network requirement to improve the training section. ANN has many forms like BPNN, MLP, and feed forward neural network (FFNN), etc. This research adopts the FFNN model (also called MLP) for risk prediction and forecasting (Erdogan and Göksu, 2014).

Before building the model, the dataset was divided into training and testing parts. Figure 2 shows the structure of a neural network with only one input layer (13 input values) and one output layer. It also has one hidden layer with eight hidden neurons and two different transfer functions (sigmoid and linear). This network runs for 1,000

epochs with the Levenberg-Marquardt training algorithm. The whole summary of neural networks is shown in Table 4.

Figure 2 Structure of neural network (see online version for colours)

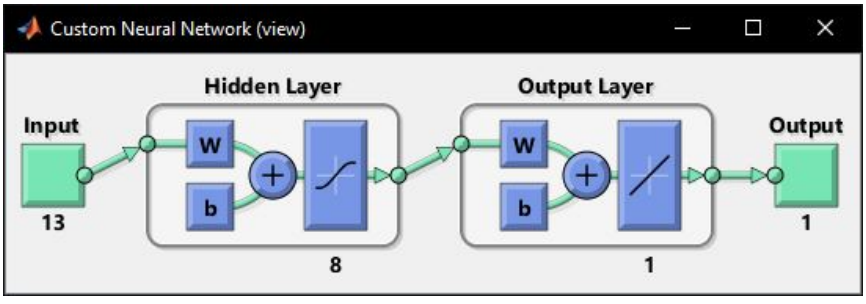


Table 4 Summary of NN structure

Characteristics	*Neural network*
Epoch	1,000
Learning algorithm	trainlm
Transfer functions	logsig, purelin
Output neurons	1
Hidden neurons	8
Input neurons	1
Training ratio	96/100
Validation ratio	2/100
Testing ratio	2/100
Number of validation checks	6

2.5.3 Bootstrap aggregated regression trees

BART is an ensemble of the decision tree, used for both classification and regression analysis. This method is frequently used for data mining to track the pattern from data. In these trees, all leaves represent the target data and all branches denote the set of input values that leads to these targets. BART is a supervised learning technique because it takes historical data to measure the target (Voyant et al., 2018). Moreover, it adopts boosting techniques to help the tree in pruning successfully and enhances its prediction accuracy. Figure 3 shows the basic structure of BART. In the first step, it takes the n out of n sample original dataset, trains each tree in the ensemble individually, and creates new training sets. At the end, all prediction results are combined to get the final result (Harrou et al., 2019; Rao, 2000). In this study, TreeBagger property is used to evaluate BART model to reduce the overfitting problem and improves the generalisation. It uses the random forest algorithm to select the random set of predictors to grow and splits multiple decision trees. Table 5 shows the properties of BART use in this research.

Figure 3 Basic idea of BART (see online version for colours)

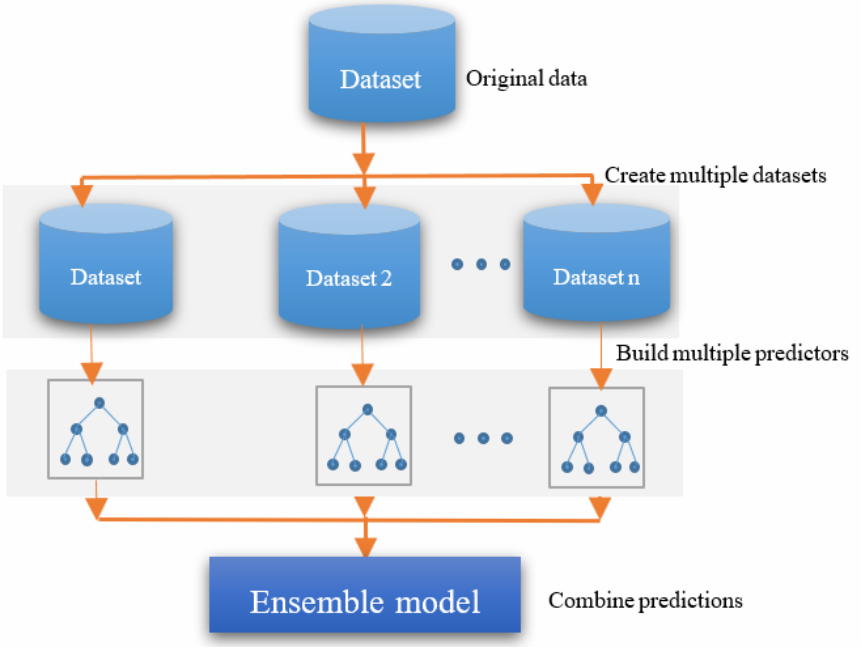


Table 5 Characteristics of BART

Characteristics	*BART*
Function name	Treebagger
Numtrees	20
Method	Regression
Minleaf	20

2.5.4 Radial basis neural network

RBNN is the most widely used model which provides the best results in different applications like time series prediction, curve fitting, function approximation, and classification problems (Shen et al., 2011). It is a three-layered feed- forward network having an input layer, an output, and hidden layers. Its first two layers can be single or multiple depending on the requirement of the model (Dash and Dash, 2015; Li et al., 2017).

Figure 4 shows the structure of RBNN used in this research. This model consists of one input layer with 13 inputs, two hidden layers with two different transfer functions and one output layer to calculate output in response to an input. The function newrbe is used to design RBNN model structure. In newrbe function, the spread is used to smooth the function approximation. Taking the too large value of smooth can cause numerical problems. The full summary of RBNN is shown in Table 6.

Figure 4 RBN structure (see online version for colours)

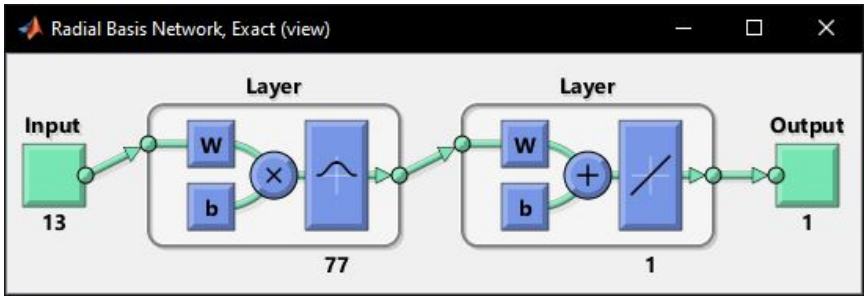


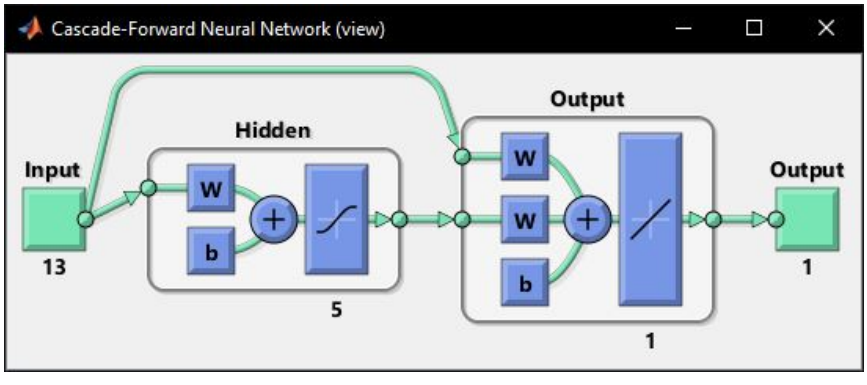
Table 6 Summary of RBNN

Characteristics	*Radial basis neural network*
Input layers	1
Number of inputs	13
Hidden layers	1
Weighted input	dist, dotprod
Net input	netprod
Spread	0.09

2.5.5 Cascade-forward neural network

CFNN is a type of FFNN in which input connects to every previous layer. The structure of CFNN is shown in Figure 5. It takes two arguments: hiddensizes and trainfunction. Hiddensizes set the size of hidden layers and trainfunction provides algorithms like trainlm to train the network (Amjady and Keynia, 2009; Mehmood and El-Hawary, 2014; Moshkbar-Bakhshayesh, 2019). The characteristics of CFNN are shown in Table 7.

Figure 5 CFNN structure (see online version for colours)



2.5.6 Implementation of models

This research goal was to forecast four different targets under four different scenarios that are explained above. All models were configured and each situation was evaluated

with four performance metrics mean absolute percentage error (MAPE), mean absolute error (MAE), daily peak MAPE and root mean square error (RMSE).

Table 7 Characteristics of CFNN

Characteristics	*Cascade forward neural network*
Input layers	1
Number of inputs	13
Hidden layers	1
Hidden neurons	5
Transfer function	logsig, purelin
Output layers	1
Number of output	1

2.6 Analysis and interpretation

The forecasting results should be analysed and compared with actual results to calculate the accuracy of the model. It improves the reliability and validity of classifiers. There were two main objectives of this phase.

- figure out the best scenario in which a model provides good results
- compare all models with their best scenarios to get the best forecaster.

2.7 Result validation

At this stage, all models are compared with each other with different performance metrics and in case of invalidation, the model is modified and again starts processing for risk analysis. This study used four different performance metrics to validate results: MAE, MAPE, daily peak MAPE and RMSE (Abdullah, 2014; Noor et al., 2008).

$$MAPE = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{\widetilde{y}_t - y_t}{y_t}\right| \tag{5}$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\widetilde{y}_t - y_t| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\widetilde{y}_t - y_t)^2} \tag{7}$$

$$DailyPeakMAPE = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{\max(\widetilde{y}_t) - \max(y_t)}{\max(y_t)}\right| \tag{8}$$

Here, $\widetilde{y}_t$ is the forecast value, y_t is the actual value, $t = 1$, and $t = T$ for the forecast period. RMSE and MAE are used for the comparison between actual and forecasted values because these are scale-dependent accuracy measures. MAPE is the conversion of MAE in a percentage format. Daily peak MAPE calculated the interest rate on bond for

each day. While evaluating the accuracy rate of forecasting models, the lower values of these metrics showed a higher rate of forecasting accuracy. These performance metrics are used in each scenario to predict the best forecasting event for 3, 5, 10, and 20 years targets.

Table 8 Best forecasting results of each model according to features

Model	*Target (years)*	*Feature*	*MAPE (%)*	*MAE (std dev)*	*Daily peak MAPE (%)*	*RMSE (std dev)*
NN	3	Generated	19.5	2.47	17.3	3.37
	5	Generated	3.6	0.43	22.1	0.67
	10	Generated	12.9	1.55	0.6	3.07
	20	LASSO	23.3	0.23	42.9	0.25
BART	3	Generated	18.3	2.31	0.1	2.35
	5	Generated	15.6	1.92	0.1	2.13
	10	Generated	12.3	1.48	0.1	1.55
	20	LASSO	56.1	0.56	47.4	0.56
RBNN	3	LASSO	2.0	0.25	1.4	0.36
	5	Generated	1.3	0.16	0.9	0.33
	10	Generated	7.0	0.84	5.9	0.87
	20	LASSO	0.0	0.00	0.0	0.00
CFNN	3	LASSO	12.8	1.61	27.8	1.95
	5	Generated	11.5	1.41	57.2	1.88
	10	Generated	18.6	2.22	23.0	3.20
	20	LASSO	20.0	0.20	37.8	0.26

2.8 Result report

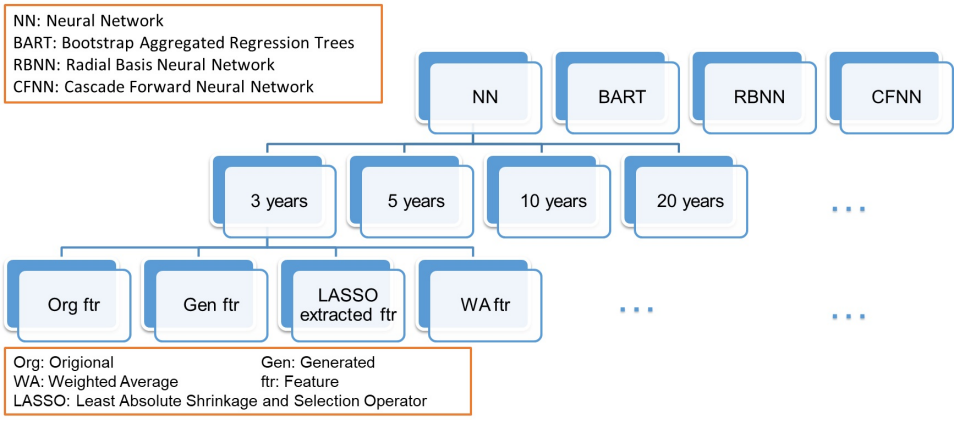
In the last phase all results are summarised and findings are calculated. The best results of all models with features are shown in Table 8.

3 Results and discussion

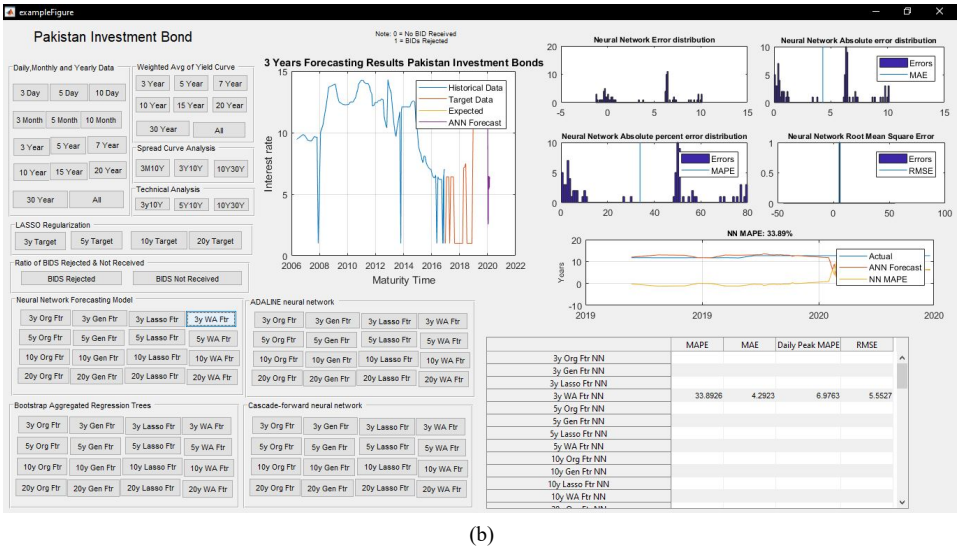
Figures 6(a) and 6(b) illustrate the logical and graphical view of the forecasting system used in this research. Each model was used to forecast each target and every single target was run under four scenarios. It means each model produced 16 results and the total number of outcomes generated by all models was 64. Table 4 only elaborates the best results of each model along features where generated and LASSO were reported as best scenarios. The graphical view of above table is shown in Figure 7(a) comparing the models along their features in terms of MAPE, similarly Figures 7(b), 7(c), 7(d) are comparing models in terms of MAE, Daily Peak MAPE and RMSE. In each figure it can be seen that RBNN has a lower error rate than other models while others are providing worse results. Previous studies used RBNN in different time series applications and highlighted numerous benefits (Dash and Dash, 2015; Li et al., 2017; Shen et al., 2011). In fact, this study compared it with ANN, adopted from literature, and concluded the same benefits mentioned in the literature. To increase the justification

of the obtained results see Figure 8: historical and target data were used for training and testing the models, while expected values were original data used for the comparison with forecasted values. It can be seen in Figures 8(a), 8(b) and 8(c), the forecasted and expected values are very closely related with each other. Whereas Figure 8(d) shows a very good fit for both data values, and error rate of 20 years target became 0.0 in all performance metrics. In summary, the main contributions of this research are to: explore more ML algorithms for yield curve forecasting, provide a framework to handle financial big data, identify the feature sets under which this methodology works well.

Figure 6 Presentation of the forecasting application, (a) logical view (b) graphical view (see online version for colours)

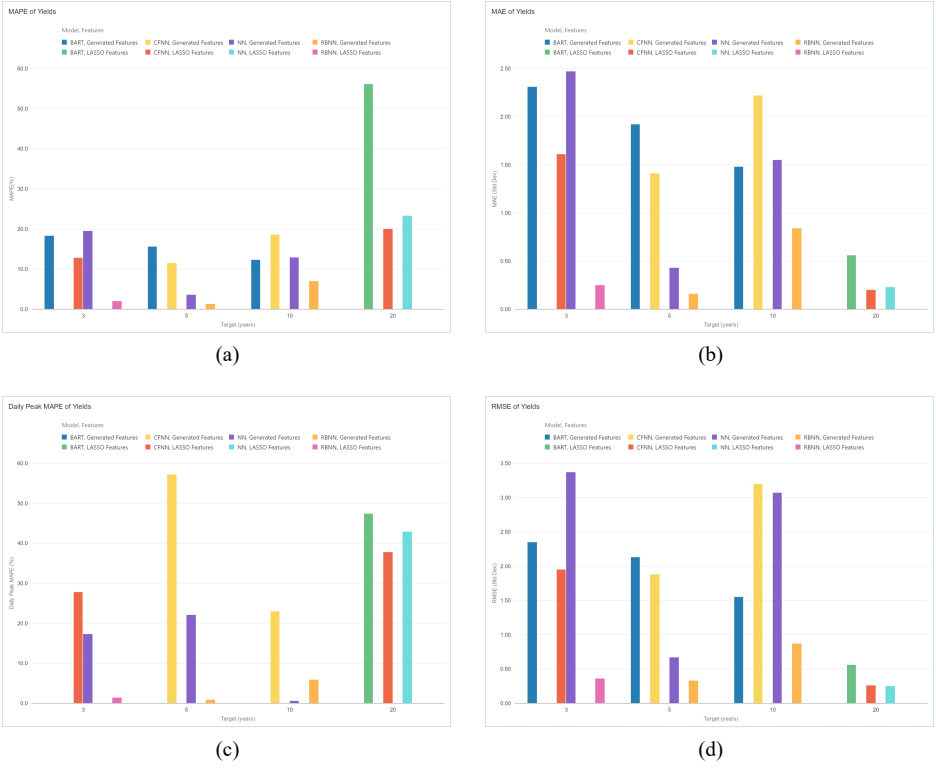


(a)



(b)

Figure 7 Comparison of models along features, (a) MAPE of yields (b) MAE of yields (c) daily peak MAPE of yields (d) RMSE of yields (see online version for colours)



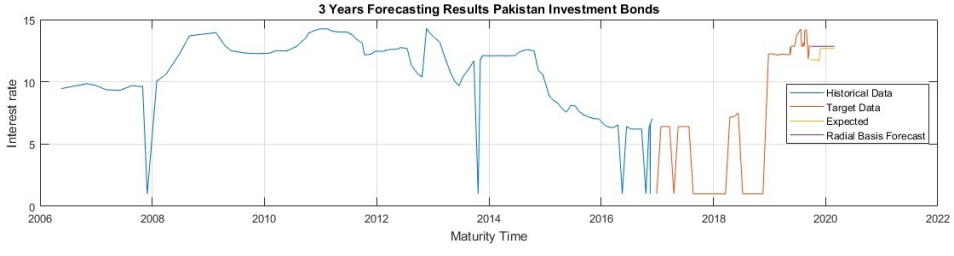
(a) (b)

(c) (d)

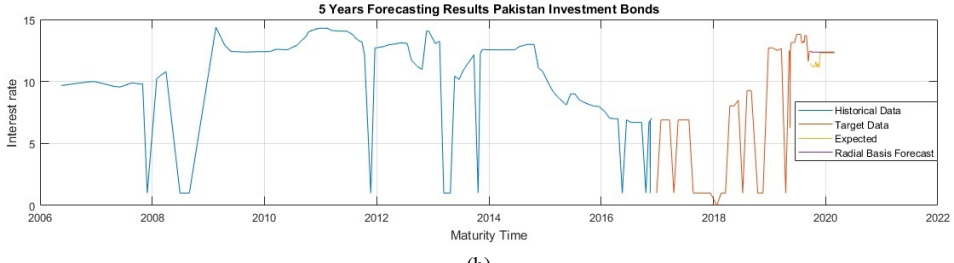
4 Conclusions

In the research, a very large and innovative topic 'risk' has been discussed in the field of computer science. In the technological era, as everything is going to be automated investigators are also trying to resolve the business issues automatically. This study focuses on interest rate risk, with the help of machine learning algorithms. Moreover, a framework is suggested to enhance the forecasting results using both big data and ML techniques. The main advantages of this research are: improve big data handling using framework, better forecasting by exploring models and increase their flexibility with multiple scenarios. Future recommendations are to explore further ML models to improve the forecasting accuracy rate using multitask learning techniques. Also, generate further new features, variables and identify scenarios under which this methodology can be used to improve performance. Additionally, more research will be needed to justify the framework scope for risk analysis.

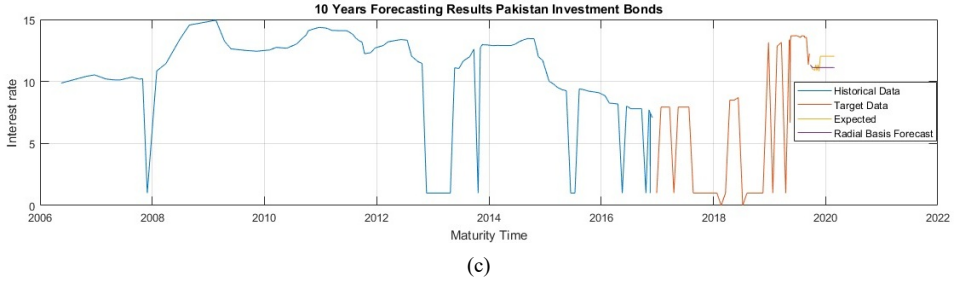
Figure 8 Final RBNN forecasting results for interest rate risk, (a) 3 years result with LASSO features (b) 5 years result with LASSO features (c) 10 years result with LASSO features (d) 20 years result with LASSO features (see online version for colours)



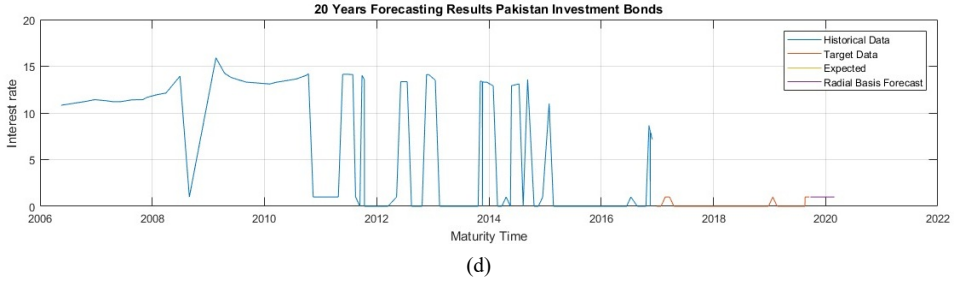
(a)



(b)



(c)



(d)

Acknowledgements

The authors would like to thank Dr. Muhammad Mohsin Nazir, an Associate Professor of the Computer Science Department, Dr. Ayesha Iqbal and Dr. Ayesha Afzaal for their support.

References

Abdullah, M.M.A.B. (2014) 'Filling missing data using interpolation methods: study on the effect of fitting distribution', *Key Engineering Materials*, Vols. 594–595, pp.889–895.

Amjady, N. and Keynia, F. (2009) 'Day-ahead price forecasting of electricity markets by a new feature selection algorithm and cascaded neural network technique', *Energy Conversion and Management*, Vol. 50, No. 12, pp.2976–2982.

Assefi, M., Behravesh, E., Liu, G. and Tafti, A.P. (2017) 'Big data machine learning using Apache Spark MLlib', *2017 IEEE International Conference on Big Data (Big Data)*, pp.3492–3498.

Bharill, N., Tiwari, A. and Malviya, A. (2016) 'Fuzzy based scalable clustering algorithms for handling big data using Apache Spark', *IEEE Transactions on Big Data*, Vol. 2, No. 4, pp.339–352.

Boyacioglu, M.A., Kara, Y. and Baykan, Ö.K. (2009) 'Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: a comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey', *Expert Systems with Applications*, Vol. 36, No. 2, Part 2, pp.3355–3366.

Brandimarte, P. (2006) *Numerical Methods in Finance and Economics: A MATLAB-Based Introduction*, 2nd ed., Wiley-Interscience.

Carbonneau, R., Laframboise, K. and Vahidov, R. (2008) 'Application of machine learning techniques for supply chain demand forecasting', *European Journal of Operational Research*, Vol. 184, No. 3, pp.1140–1154.

Dash, R. and Dash, P.K. (2015) 'A comparative study of radial basis function network with different basis functions for stock trend prediction', *2015 IEEE Power, Communication and Information Technology Conference (PCITC)*, pp.430–435.

Dibike, Y.B. and Solomatine, D.P. (2001) 'River flow forecasting using artificial neural networks', *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, Vol. 26, No. 1, pp.1–7.

Enke, D. and Mehdiyev, N. (2013) 'Type-2 fuzzy clustering and a type-2 fuzzy inference neural network for the prediction of short-term interest rates', *Procedia Computer Science*, Vol. 20, pp.115–120.

Erdogan, O. and Göksu, A. (2014) 'Forecasting euro and Turkish lira exchange rates with artificial neural networks (ANN)', *International Journal of Academic Research in Accounting, Finance and Management Sciences*, Vol. 4, No. 4, pp.307–316.

Harrou, F., Saidi, A. and Sun, Y. (2019) 'Wind power prediction using bootstrap aggregating trees approach to enabling sustainable wind power integration in a smart grid', *Energy Conversion and Management*, Vol. 201, p.112077.

Ince, H. and Aktan, B. (2009) 'A comparison of data mining techniques for credit scoring in banking: a managerial perspective', *Energy Conversion and Management*, Vol. 10, No. 3, pp.1611–1699.

Joseph, A., Larrain, M. and Singh, E. (2011) 'Predictive ability of the interest rate spread using neural networks', *Procedia Computer Science*, Vol. 6, pp.207–212.

Kanevski, M., Maignan, M., Pozdnoukhov, A. and Timonin, V. (2008) 'Interest rates mapping', *Physica A: Statistical Mechanics and its Applications*, Vol. 387, No. 15, pp.3897–3903.

Kim, W. (2019) 'Government spending policy uncertainty and economic activity', *Journal of Macroeconomics*, Vol. 61, No. 4, pp.103–124.

Kumar, R. (2014) *Strategies of Banks and Other Financial Institutions*, Academic Press, San Diego.

Li, Y., Wang, X., Sun, S., Ma, X. and Lu, G. (2017) 'Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks', *Transportation Research Part C: Emerging Technologies*, Vol. 77, pp.306–328.

Mehmood, S.T. and El-Hawary, M. (2014) 'Performance evaluation of new and advanced neural networks for short term load forecasting', *2014 IEEE Electrical Power and Energy Conference*, pp.202–207.

Moshkbar-Bakhshayesh, K. (2019) 'Comparative study of application of different supervised learning methods in forecasting future states of npps operating parameters', *Annals of Nuclear Energy*, Vol. 132, pp.87–99.

Noor, N.M., Yahaya, A.S. and Abdullah, M.M.A.B. (2008) 'Estimation of missing values in air pollution data using single imputation techniques', *Key Engineering Materials*, Vol. 34, No. 3, pp.341–345.

Nunes, M., Gerding, E., McGroarty, F. and Niranjan, M. (2019) 'A comparison of multitask and single task learning with artificial neural networks for yield curve forecasting', *Expert Systems with Applications*, Vol. 119, pp.362–375.

Oh, K.J. and Han, I. (2000) 'Using change-point detection to support artificial neural networks for interest rates forecasting', *Expert Systems with Applications*, Vol. 19, No. 2, pp.105–115.

Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., Benítez, J.M. and Herrera, F. (2017) 'Nearest neighbor classification for high-speed big data streams using Spark', *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 47, No. 10, pp.2727–2739.

Rao, J.S. (2000) 'Bootstrapping to assess and improve atmospheric prediction models', *Data Mining and Knowledge Discovery*, Vol. 4, No. 1, pp.29–41.

Sakr, S., Maamar, Z., Awad, A., Benatallah, B. and Aalst, W.M.P.V.D. (2018) 'Business process analytics and big data systems: a roadmap to bridge the gap', *IEEE Access*, Vol. 6, pp.77308–77320.

Sambasivan, R. and Das, S. (2017) 'A statistical machine learning approach to yield curve forecasting', *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)*, pp.1–6.

SBP (1956) *State Bank of Pakistan* https://www.sbp.org.pk/index.asp (accessed 4 October 2020).

Shen, W., Guo, X., Wu, C. and Wu, D. (2011) 'Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm', *Knowledge-Based Systems*, Vol. 24, No. 3, pp.378–385.

Suleiman, D., Al-Zewairi, M. and Naymat, G. (2017) 'An empirical evaluation of intelligent machine learning algorithms under big data processing systems', *Procedia Computer Science*, Vol. 113, pp.539–544.

Sun, Z-L., Choi, T-M., Au, K-F. and Yu, Y. (2008) 'Sales forecasting using extreme learning machine with applications in fashion retailing', *Decision Support Systems*, Vol. 46, No. 1, pp.411–419.

Tavana, M., Abtahi, A-R., Caprio, D.D. and Poortarigh, M. (2018) 'An artificial neural network and bayesian network model for liquidity risk assessment in banking', *Neurocomputing*, Vol. 275, pp.2525–2554.

Tripathy, P., Rautaray, S.S. and Pandey, M. (2017) 'Map-reduce based parallel support vector machine for risk analysis', *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp.300–303.

Voyant, C., Motte, F., Notton, G., Fouilloy, A., Nivet, M-L. and Duchaud, J-L. (2018) 'Prediction intervals for global solar irradiation forecasting using regression trees methods', *Renewable Energy*, Vol. 126, No. C, pp.332–340.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M-L., Paoli, C., Motte, F. and Fouilloy, A. (2017) 'Machine learning methods for solar radiation forecasting: a review', *Renewable Energy*, Vol. 105, No. C, pp.569–582.

Wackernagel, H. (2003) 'Geostatistical models and kriging', *IFAC Proceedings Volumes*, Vol. 36, No. 16, pp.543–548.

Zhang, W. (2017) 'Machine learning approaches to predicting company bankruptcy', *Journal of Financial Risk Management*, Vol. 6, No. 4, pp.364–374.