Decision Making: Applications in Management and Engineering
Vol. 4, Issue 2, 2021, pp. 225-240.
ISSN: 2560-6018
eISSN: 2620-0104
DOI: https://doi.org/10.31181/dmame210402215s

A NOVEL PREDICTION ALGORITHM FOR MULTIVARIATE DATA SETS

Pinki Sagar[1], Prinima Gupta [1] and Rohit Tanwar [2*]

[1] Computer Science and Technology, Manav Rachna University, Haryana, India
[2] School of Computer Science, University of Petroleum & Energy Studies, Dehradun Uttarakhand, India

Received: 26 April 2021;
Accepted: 14 July 2021;
Available online: 15 July 2021.

Original scientific paper

Abstract: *Regression analysis is a statistical technique that is most commonly used for forecasting. Data sets are becoming very large due to continuous transactions in today's high-paced world. The data is difficult to manage and interpret. All the independent variables can't be considered for the prediction because it costs high for maintenance of the data set. A novel algorithm for prediction has been implemented in this paper. Its emphasis is on the extraction of efficient independent variables from various variables of the data set. The selection of variables is based on Mean Square Errors (MSE) as well as on the coefficient of determination r^2p, after that, the final prediction equation for the algorithm is framed based on of deviation of the actual mean. This is a statistical-based prediction algorithm that is used to evaluate the prediction based on four parameters: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and residuals. This algorithm has been implemented for a multivariate data set with low maintenance costs, preprocessing costs, lower root mean square error and residuals. For one dimensional, two-dimensional, frequent stream data, time series data and continuous data, the proposed prediction algorithm can also be used. The impact of this algorithm is to enhance the accuracy rate of forecasting and minimized the average error rate.*

Keywords: *Coefficient of determination, Mean square error, Actual means, Multiple Linear Regression (MLR), Root Mean Square Error (RMSE), Mean Square Error (MSE).*

1. Introduction

Regression techniques come under the category of supervised learning methods in which the existing training data sets can be used as guidance and to supervise the

E-mail addresses: pinki.fet@mriu.edu.in (P. Sagar), prinima@mru.edu.in (P. Gupta), rohit.tanwar.cse@gmail.com (R. Tanwar).

complete learning and prediction process. The results of supervised learning approaches are dependent on algorithms and their complexity. In the regression techniques, new values are predicted for future analysis, which will be calculated based on historical or previous data sets. Linear Regression fits a straight line and it has two components b_0 (intercept) and coefficient b_1 and one predictor termed as the independent variable. In today's scenarios, data sets are maintained with multiple attributes and it requires so much processing time and costs for prediction. The cost of preprocessing and maintenance is depending on the type of data sets but at the time of analysis, it is not necessary to consider all attributes. In this paper prediction algorithm is introduced based on actual means which improve the prediction rate and reduce the cost of maintenance of the data. The regression line includes the following properties: The line restricts the aggregate of squared differentiation between observed values (y dependent variable) and foreseen characteristics (the $\hat{y}$ values enrolled from the regression line). The regression line experiences the mean of the 'x' and mean of the 'y'. In the linear regression, the b_0 is considered as the intercept of the regression equation and it is the incline of the regression line. The regression coefficient b_1 is the average change in the dependent variable 'y' for a per-unit change in the independent variable 'y'.

1.1. Regression Coefficients

Regression coefficients are estimates of the unknown population parameters and describe the relationship between a dependent and the independent variable. These are identified using methods, such as least square and matrix form. The Least square method is a modest linear forecasting approach, in which there is only a binary dependent variable and the other one is a neutral or independent variable. The equation for prediction using regression is:

$$\hat{y} = b_0 + x * b_1 \tag{1}$$

Least-squares regression coefficients Daniya et al. (2020):

$$b_1 = \frac{\sum[(x_{(i-n)} - \bar{x})((y_{(i-n)} - \bar{y}))]}{\sum[((x_{(i-n)} - \bar{x}^2)} \tag{2}$$

$$b_0 = y - x * b_1 \tag{3}$$

In equation (1) $\hat{y}$ is the projected value of the reliant variable in linear regression two variables are used b_0 and b_1. The $x(i\text{---}n)$ is the value of the independent variable for observation or new predicted values. For observed values $y(i\text{---}n)$ is used, in which y is a dependent variable. x implies x score, and y is the mean y score. Equation (2) and (3) Daniya et al. (2020) represents the method of coefficients calculation. With Multiple linear regression, things are getting progressively jumbled and confused. In Multiple linear regression, 'n' free factors and 'n + 1' relapse coefficients and ordinary conditions are used. Finding the least-squares arrangement includes solving 'n + 1' condition with 'n + 1' questions. Equation (1) is eligible only for prediction in the one-dimensional data set. If the prediction is done in a multivariate data set then there will be many independent variables but all are not required for prediction. So, the proposed algorithm will work on the selection of attributes that have the highest weightage and more suitable for prediction, after selection of variable prediction equation will be formed based on the actual mean.

1.2. Research Contribution

In this paper, an algorithm (MIPA) is explained that will be applicable for multivariate data sets. In the preprocessing part, irrelevant variables are reduced. Based on the selected variables actual mean has been calculated for identifying the coefficients. The selection of variables is based on the coefficient of determination (r^2p) and mean square errors (MSE). This algorithm can be applied on various types of data set and reduced the errors like RMSE, MAE, and MAPE so that the accuracy rate of prediction can be improved.

2. Related Work

Regression techniques are important tools for prediction and analysis. It indicates the significant associations between the dependent variable and independent variable and the strength of the impact of multiple independent variables on a dependent variable. Chai et al. (2007) introduced two prediction algorithms that are applicable for one-dimensional and two-dimensional stream data: Frequent Item Prediction Method (FIPM) and Frequent Temporal Pattern Data Stream (FTPDS). Stream data converted into discrete data to get dependent and independent variables so that regression models can be applied. In these algorithms, there were some limitations that were recovered by Sequence Forecast Algorithm Plane Regression (SFAPR) introduced. This plane regression algorithm is based on linear regression for two-dimensional data sets and reduces the error rate. Kavitha et al. (2016) discussed that, after the advancement of technologies in big data, data analytic has been developed wonderfully in today's environment. The measurable strategies are utilized for the assessment of prescient models; the choice of accurate systems depends on the prerequisites of the information. The expectation and determining are done generally with time-series data sets. The majority of the applications of prediction are: climate determining, account and securities exchange join recorded information with the present gushing information for better exactness. In this paper, the author divided the time arrangement information using a regression model. Linear and multiple linear regression models are connected using the training data set for applying and also for preparation of informational assortment so that it can operate the right model for improvement.

Ostertagova et al. (2016) presented the application of linear regression algorithm for processing of stress state data which were collected through drilling into a Harmonic Star Method (HSM) it was used for the collection of final data. The non-commercial software based on the harmonic star method enables us to automate the process of measurement for the direct collection of experiment data. Such programming empowered us to gauge worries in a specific purpose of the analyzed surface and, simultaneously, separate these anxieties. For example, a camera was utilized to move the image of its chromatic edges legitimately to a computer.

Mustapha & Fadzil (2015) presented a regression algorithm for vendors to forecast their yearly profit and it is based on their historical data. Using a forecasting approach vendor can prepare their evaluation exercise. In this article, the author used various regression techniques that analyze the vendor's performance. The performance report demonstrates the capability of data mining tasks in helping the Entrepreneur Development Unit (EDU) to predict vendors' performance and to identify groups of on performance and under-performance. The Entrepreneur Development Unit was responsible for managing a big group of vendors that hold contracts with the company.

Khan et al. (2016) discussed a non-linear regression by assuming that the data depend on a variety of folds. They divided the data space into multiple areas to construct a partitioned linear regression analysis as an estimation of the non-linearity among the experiential and the expected data, in place of setting up the range and limitations of the particular category, the Algorithm was exposed to immediately adapt to the differences in the data and it was very successful for high dimensional data as well as for small data sets. Saptawati et al. (2015) stated that the major activity in the mineral industry was most significant and costly in drilling. Although finding targets for drilling, geologists were using qualitative study, which resulted in a lot of failures in drilling. The authors worked on the analysis of methods, used for mining, and used to retrieve the facts in the outcome, the categorization of informed data, and mining of common item sets that can maintain the forecasting of drilling targets. The objective of the work is to reduce the threat of failure of drilling and hold the industry's decision and decide on a new target in drilling. Ilayaraja & Meyyappan (2015) discussed the data mining techniques and applied them in many areas of medicine for different objectives. They partitioned a process to estimate the risk factors of the patients who were having symptoms of heart disease through the collected frequent data sets. Data sets for the heart patients were collected from the medical institutes or hospitals. Frequent data item sets are produced and depend on the selected symptoms and minimum support value. The frequent or common data sets which were extracted could help the doctors to make decisions in diagnostics and to predict the level of risk at an early stage so that immediate treatment could be provided. The projected approach could be applied to a data set of medical fields which helps in predicting the factors that affect risk with the level of risk and the patients based on selective factors item sets.

Yang et al. (2019) used both a linear model and a nonlinear model to predict the future cash flow. A hybrid model integrating linear and nonlinear was constructed to enhance the prediction effect and calculate the fund reserve ratio and improved the accuracy rate of prediction. In the literature survey, it has been discussed that the prediction algorithm discussed by Zhao & Li (2005) is used for two-dimensional stream data that was based on plane regression. Chai et al. (2007) discussed the prediction algorithm for one-dimensional and two-dimensional stream data. Later on, a non-linear regression algorithm has been proposed for two-dimensional and one-dimensional stream data. After those algorithms for multi-dimensional data sets were discussed but it takes a lot of cost and time for the maintenance and preprocessing of data. In this paper, we minimize the cost of storing and preprocessing data sets and increase the accuracy rate of prediction and decrease the error rate via the proposed method.

Antoniadis et al. (2021) reviewed the sector of sensitivity analysis and targeted the link between random forest and global sensitivity analysis (GSA). The concept is to use the random forest technique as an effective non-parametric method for building a meta-model that permits effective sensitivity analysis. In addition to its straightforward relevance to regression problems, the random forest methods additionally have the flexibility to implicitly handle correlation and high-dimensional data. Authors have used the rank-based random forest (RF) variable index to define sensitivity indexes. The author further reviewed the acceptable tool set for quantifying the importance of variables and used these tools to cut back the spatial property of the model, thereby conducting sensitivity analysis studies that might not be performed.

Yıldırım et al. (2021) used a preferred deep learning tool known as long short-run memory (LSTM), which has been shown to be terribly effective in several time-series prognostication problems. They projected the hybrid model using two data sets that

mix two separate LSTMs to improve the prediction, and it was found that the model gave good results for real data.

Mukherjee et al. (2019) proposed a model to predict the images based on spatio temporal sequence forecasting problems. They trained the Convolutional Long Short Term Memory (Conv-LSTM) to learn the temporal relationships while preserving the spatial data and present in the Latent space. In this method first, the encoder and decoder network are trained to learn the spatial features of the data. After that, the Conv-LSTM is inserted between the encoder and the decoder. The weights of the encoder-decoder are freezed and then the Conv-LSTM is trained. In the experiment loss function is used to predict the next set of frames for a given set of frames in a video. Gauba et al. (2017) proposed a novel approach to predict the rating of video advertisements based on a multimodal framework combining physiological analysis of the user and global sentiment-rating available on the internet. In the framework, they record the EEG signals while the users were asked to watch the video advertisement simultaneously. To predict the rating of an advertisement using EEG data, they used the regression technique based on Random forest and then EEG-based rating is combined with NLP-based sentiment score to improve the overall prediction.

3. Proposed Work

In the literature survey, it has been identified that various algorithms have been introduced by authors for prediction using regression. Some existing algorithms and methods discussed in section 2 are used for the prediction process for forecasting. The proposed algorithm is based on the selection of efficient variables and prediction of new dependent variables with low residuals, Root Mean Square Errors, Mean Absolute Error, and Mean Absolute Percentage Error.

3.1. Problem Formulation

Forecasting is the estimation of a dependent variable 'y', based on the independent variable 'x'. Some algorithms are implemented for one-dimensional and two-dimensional data sets. These data sets can be stream, continuous or discrete data. Most of the data sets have multiple independent variables and in the prediction equation, all variables are used for prediction, which may cause the extra cost of maintenance of the data set, more execution time, and low accuracy rate of prediction. The proposed algorithm focused on the selection of relevant variable (independent) from multivariate data sets and improving accuracy rates because multivariate data set consist of various independent variable but all are not required for prediction model and it is very difficult to maintain huge data set with multiple variables due to high cost of maintenance and pre-processing of data set. In the literature survey, it has also been found that existing algorithms are restricted to the number of independent variables for one- and two-dimensional stream data. The objective of the proposed algorithm is to reduce the errors with more accuracy in prediction. The proposed algorithm is used for data sets that have numerous independent variables and reduce the cost of maintenance by selecting the appropriate variables for prediction then the prediction equation is framed based on assumed means. The selection of variables is based on the coefficient of determination and the prediction equation is based on actual means.

This algorithm is applied on "Energy Data Prediction", from the UCI repository, (https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction). In the proposed prediction algorithm, there is one dependent variable and 'n' independent variables. The dependent variable is Humidity (Average of all areas of the building) and the independent variables are Average temperature (average of all areas of the building), Pressure, RH_out, Wind speed, visibility, etc.

3.2. Multivariate Item Prediction Algorithm (MIPA)

In the algorithm coefficients are calculated for independent variables using the actual mean of dependent and independent variables, in Step 4 formula for coefficient calculations has been explained.

Algorithm Name: Multivariate Item Prediction Algorithm (MIPA)

Input: Multivariate data set
Output: Low Residuals, RMSE, MAE, MAPE during the prediction

Step 1: Take 2^n regression equations 'n' is the number of independent variables.

Step 2: Categorize all the equations into Models, Model 1= no independent variable, Model 2 = 1 independent variable, Model 3 = 2 independent variables and so on.

Step 3: Using ANOVA compare r^2p and MSE for each regression model in each Model.
If ($r^2p \uparrow$ MSE $\downarrow$) /*select the regression model if this condition is true highest value of r^2p and lowest value of MSE */
For i=0 to 2^n /*select the appropriate regression model which has the highest r^2p and lowest MSE out of 2^n regression model*/

Step 4: Select the regression model from each Model which have highest r^2p and lowest MSE.

Step 5: Compare all selected regression equation from each Model consider the regression model with lowest MSE and highest r^2p /*selection of independent variables.*/

Step 6: Find the deviation of actual mean for each selected independent variable.

$$dx(i = 1 \ldots . n) = x - \bar{x}$$

$$dy(i = 1 \ldots . n) = y - \bar{y}$$

$$byx_{(i=1\ to\ n)} = \frac{n\sum\left[(dx_{(i=1\ to\ n)} * dy) - \sum(dy_{(i=1\ to\ n)} * dy)\right]}{n\sum(dx_{(i=1\ to\ n)})^2 - (\sum(dx_{(i=1\ to\ n)}))^2}$$

Step 7: Put value of $byx_{(i=1\ to\ n)}$ in the prediction model

$$\hat{Y} = byx_1\,(x_1 - \bar{x}_1) + byx_2\,(x_2 - \bar{x}_2) + byx_3\,(x_3 - \bar{x}_3) + \cdots + byx_n\,(x - \bar{x}) + \bar{y}$$

Step 8: Analysis of prediction algorithm (RMSE and residuals)

$$RMSE = [\sum_{k=0}^{n}((Y - \hat{Y}) * (Y - \hat{Y}))/N]$$

Residuals=Actual values- Predicted values(y- $\hat{y}$)

Algorithm MIPA has been discussed in detail as follows:

Step 1: Find the 2^n regression models in which 'n' is the number of independent variables. e.g. in the case of 4 independent variables total possible equations will be 16. In Table 1, Model 1 has not considered any independent variable, for Model 2 one independent variable is considered, for Model 3 two independent variables are considered, and so on. In Table 1 all possible regression models are shown.

Table 1. The 2^n Possible Regression equations model.

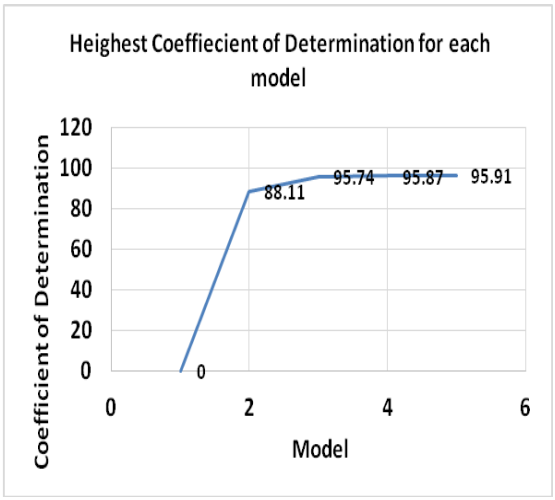
Mode1	Model 2	Model 3	Model 4	Model 5
y=b0+e	y=b0+b1X1	y= b0+b1X1+b2X2	y=b0+b1X1+b2X2+b3X3	y= b0+b1X1+b2X2 +b3X3+ b4X4
	y=b0+b2X2	y= b0+b1X1+b3X3	y=b0+b1X1+b2X2+b4X4	
	y=b0+b3X3	y= b0+b1X1+b4X4	y=b0+b1X1+b3X3+b4X4	
	y=b0+b4X4	y= b0+b2X2+b3X3	y=b0+b2X2+b3X3+b4x4	
		y=b0+b2x2+b4X4		
		y= b0+b3X3+b4X4		

Step 2: Find the Analysis of Variance (ANOVA table) of each regression model (2^n). Select the regression model from each Model which has the highest r^2p (coefficient of determination) and lowest MSE. Table 2 includes the values of r^2p and MSE from each model of Table 1.

Table 2. MSE and r^2p (Coefficient of Determination)

Model 2		Model 3		Model 4		Model 5	
r^2p (%)	MSE square	r^2p (%)	MSE square	r^2p (%)	MSE square	r^2p (%)	MSE square
85.35	0.24	95.74	0.07	95.87	0.07	95.91	0.08
12.08	1.47	90.91	0.16	95.83	0.07		
88.11	0.19	93.96	0.11	93.96	0.11		
3.7	1.61	88.14	0.21	91.33	0.16		
		76.92	0.4				
		88.35	0.2				

Step 3: Model 2 consists of 1 independent variable, Model 3 consists of 2 independent variables, and Model 3 consists of 4 independent variables, and so on. The condition must be r^2p ↑ MSE ↓ means that if the value of r^2p is increasing then the value of MSE will decrease. So that Model 4 is selected in Table 2, which has values 95.87 and 0.07 belong to the first regression model of Model 4 in Table-1. It includes X_1, X_2, and X_3 independent variables; it means that these three variables are most relevant for the prediction algorithm.



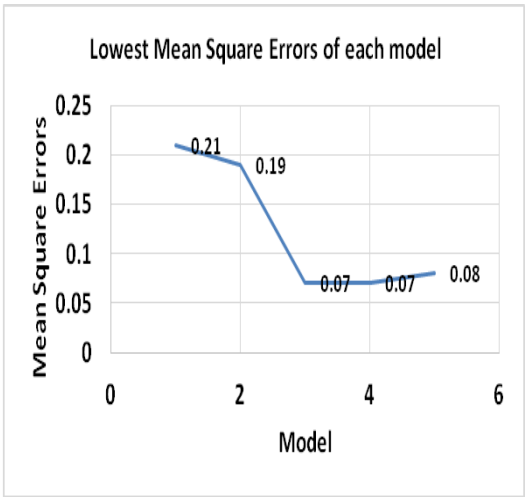


Figure 1(a). Highest coefficient of determination r^2p; **Figure 1(b).** Lowest Mean Square Error (MSE)

In Figure 1(a) values of coefficient of determination are plotted which are selected as the highest value from each model and in Figure 1(b) plotted the values of lowest MSE of Table 2. In Figure 1(a) value 95.91 is highest but MSE is high corresponding to it so that 95.87 will be selected which is corresponding to the lowest MSE i.e. 0.07.

Step 4: Derivation of proposed regression algorithm on the basis of deviation from actual mean is as follows:

Find the value of *dx* and *dy* on the basis of actual mean:

$$dx(i = 1 \ldots . n) = x - \bar{x} \quad (4)$$

$$dy(i = 1 \ldots . n) = y - \bar{y} \quad (5)$$

$$byx_{(i=1\ to\ n)} = \frac{n\sum[(dx_{(i=1\ to\ n)} * dy) - \sum(dy_{(i=1\ to\ n)} * dy)]}{n\sum(dx_{(i=1\ to\ n)})^2 - (\sum(dx_{(i=1\ to\ n)}))^2} \quad (6)$$

$$\hat{Y} = byx_1 (x_1 - \bar{x}_1) + byx_2 (x_2 - \bar{x}_2) + byx_3 (x_3 - \bar{x}_3) + \cdots + byx_n (x - \bar{x}) + \bar{y} \quad (7)$$

Step 5: Find the RMSE. Using equation (8) RMSE is calculated. If RMSE is low during the prediction means that accuracy of prediction is high.

$$RMSE = [\sum_{k=0}^{n}((Y - \hat{Y}) * (Y - \hat{Y}))N] \quad (8)$$

Step 6: Analysis of residual error (Actual y- $\hat{y}$)

In the proposed algorithm Step 1, 2 and 3 are preprocessing steps, used for the selection of efficient variables which are required for the prediction equation (7). In step 4 equations (6) represent the coefficients calculations and prediction equation (7) is drafted for forecasting of new values. As we have mentioned that this algorithm is also valid for different types of data sets, the above-explained approach was extended to a multivariate data set. Here, this method is also applied to one-dimensional data set selected from the UCI repository https://archive.ics.uci.edu/ml/datasets/Parking +Birmingham), data set contain four attributes parking id (System Code Number), the capacity of parking (Capacity), parking rates, and updated details. In this data set parking occupancy is an independent variable and parking rates are the dependent variable. The coefficient for independent variables is calculated using equations (4), (5), and (6) and then places

the values of byx for each independent variable in equation (7). This proposed prediction equation can also be applicable to the stream data set. For the prediction of stream data first, the stream data need to convert into a form of discrete data sets.

4. Implementation and Result

The algorithm is implemented in "R 3.3.2" version. One-dimensional and multivariate data sets are collected from an online repository.

4.1. Implementation

One dependent variable (Humidity) and 4 independent variables (Wind speed, Average, temperature, t out, press m hg) are included in multivariate data collection. In one dimensional data set four attributes parking id (System Code Number), the capacity of parking (Capacity), parking rates are available. The capacity of parking is an independent variable and parking rate is a dependent variable. For multivariate datasets, all possible equations for independent variables are considered in this process. For the prediction system, appropriate independent variables can be identified using preprocessing. In this process, 2^n equations are considered as n is the number of independent variables, and these equations are shown in Table-1. In the multivariate data sets, it is not needed to consider all variables in the prediction algorithm. Coefficient of determination and MSE are used for finding relevant variables. All independent variables have no equal significance or priority so few variables can be eliminated. For one-dimensional data set, there is no need to process the data set, only the regression equation will be applied on attributes x and y.

4.2. Results

In this paper, a MIPA algorithm is compared with MLR because it deals with multiple independent variables. RMSE values for MLR and MIPA algorithms are plotted in Figure 4. As compared with MLR, it has been evaluated that RMSE's are low for the MIPA algorithm. In Table 3 residuals are analyzed or compared which are generated through MLR and MIPA algorithms. These values are corresponding to humidity. By analyzing Table 3 it can be easily observed that the improved algorithm gives low error rates during the prediction of humidity based on temperature, wind speed, etc. independent variables.

Table 3. Residuals using MLR and MIPA

Independent variables						
Dependent variable Humidity	Average(temp)=X_1	Press_mm_hg=X_2	RH_out=X_3	Windspeed=X_4	Residuals MIPA	Residuals MLR
50.91	17.1674074	733.5	92	7	0.42107573	0.99597124
50.83	17.1496296	733.6	92	6.66666666	0.00879531	1.17010151
50.63	17.1037037	733.7	92	6.33333333	0.55354625	1.48437316
50.57	17.0670370	733.8	92	6	0.95598559	1.64734098
50.73	17.0707407	733.9	92	5.66666666	1.10794818	1.56379616
50.79	17.0485185	734	92	5.33333333	1.37981300	1.59155517
50.79	17.0407407	734.1	92	5	1.70193451	1.6768479
50.8	17.0185185	734.166666	91.8333333	5.16666666	1.63454412	1.67028356
50.9	17.0185185	734.233333	91.6666666	5.33333333	1.46779281	1.56250345
51.05	17.0396296	734.3	91.5	5.5	1.21976676	1.38058893
51.23	17.0667592	734.366666	91.3333333	5.66666666	0.93158178	1.15983686
51.47	17.1103703	734.433333	91.1666666	5.83333333	0.58419480	0.87670657
51.85	17.1851851	734.5	91	6	0.04521525	0.41176678
52.68	17.2149074	734.616666	90.5	6	0.65000026	0.44031096
53.52	17.2522222	734.733333	90	6	1.37553362	1.32006378
53.5	17.2866666	734.85	89.5	6	1.24106697	1.33981661
53.38	17.3107407	734.966666	89	6	0.98628249	1.24189434
53.38	17.3133333	735.083333	88.5	6	0.83118018	1.24629699
52.97	17.3196296	735.2	88	6	0.27623678	0.84953717
54.37	17.3748148	735.233333	87.8333333	6	1.67429658	2.2952218
55.07	17.465	735.266666	87.6666666	6	2.42625301	3.0850061
54.9	17.4588888	735.3	87.5	6	2.19335931	2.90766546

4.3. Analysis with Existing algorithms

MIPA algorithm can be used for various types of data sets such as stream data sets (one dimensional, two-dimensional), time-series data set, and multivariate data sets. It identifies the relevant variables from the data set which are the best suitable for the prediction. In FIPM, FTPDS preprocessing of data is done with the use of sliding window protocol and it is applicable only for one-dimensional and two-dimensional stream data sets. MLR is used for data sets where multiple independent variables are used for prediction, but it consumes lots of time and cost. Here in Table 4, MIPA is analyzed with MLR, FIPM, and FTPDS, based on residuals parameters. Residuals are the errors or difference values of the dependent variable and the observed values. In Table 3 independent variables are mentioned, humidity is predicted based on the selected independent variables.

Table 4. Analysis of residuals of existing algorithm with proposed algorithm.

Humidity	MIPA	MLR	FIPM	FTPDS
50.91	0.421075736	0.9959712	1.3323720	0.58600286
50.83	0.008795312	1.1701015	1.4105585	0.786391724
50.63	0.553546256	1.4843731	1.6086900	1.114464984
50.57	0.955985598	1.6473409	1.6668764	1.297415314
50.73	1.107948187	1.5637961	1.5050629	1.252681248
50.79	1.379813003	1.5915551	1.4431944	1.313070113
50.79	1.70193451	1.6768479	1.4413808	1.430897512
50.8	1.634544126	1.6702835	1.4323151	1.420897512

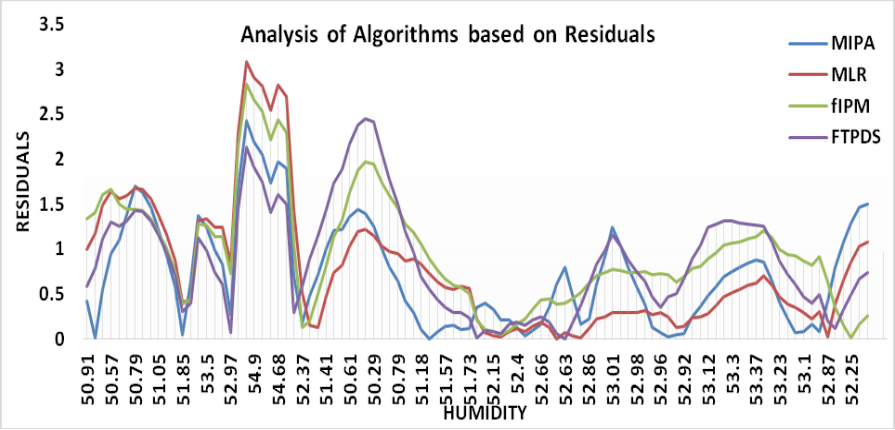


Figure 2 Plotting of residuals during prediction

In Figure 2 residuals of existing algorithms and MIPA algorithm are plotted compared to MIPA and have low residuals in comparison of MLR, FIPM, and FTPDS.

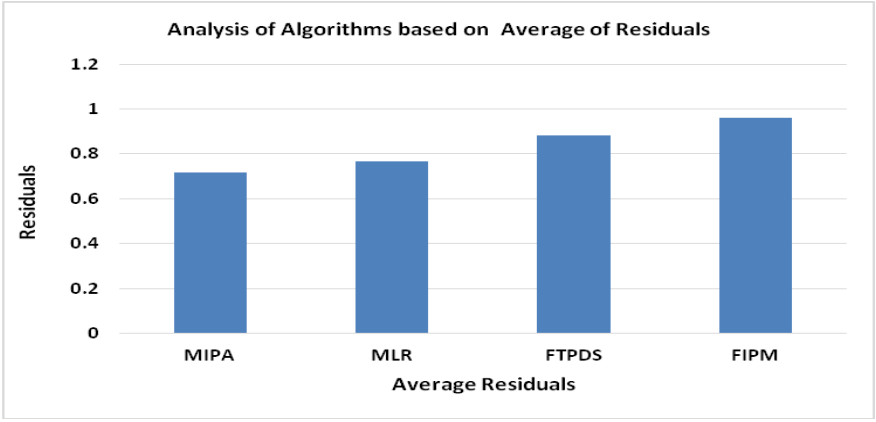


Figure 3. Analysis of average of Residuals

In Figure 3 we have compared the average residuals of algorithms MIPA has low average residual values 0.71525 and average residual values of MLR, FIPM, and FTPDS are 0.7664, 0.9615, and 0.8799 respectively.

5. Analysis of Results

The analysis of MIPA has been done on the basis of parameters such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and residuals.

5.1. Analysis of RMSE

RMSE is the standard deviation of residuals that are the prediction errors. It is a measure of how these residuals are spaced out. Residuals are a measure of how far data points are out from the regression line. By analysis of MLR and MIPA, it has been identified that MIPA has low RMSE values in comparison to MLR. In Figure 4, MIPA's RMSE values are 0.647776757, 0.582948804, 0.516691943, and 0.496299287 are respectively plotted which are low in comparison to RMSE values calculated by MLR.

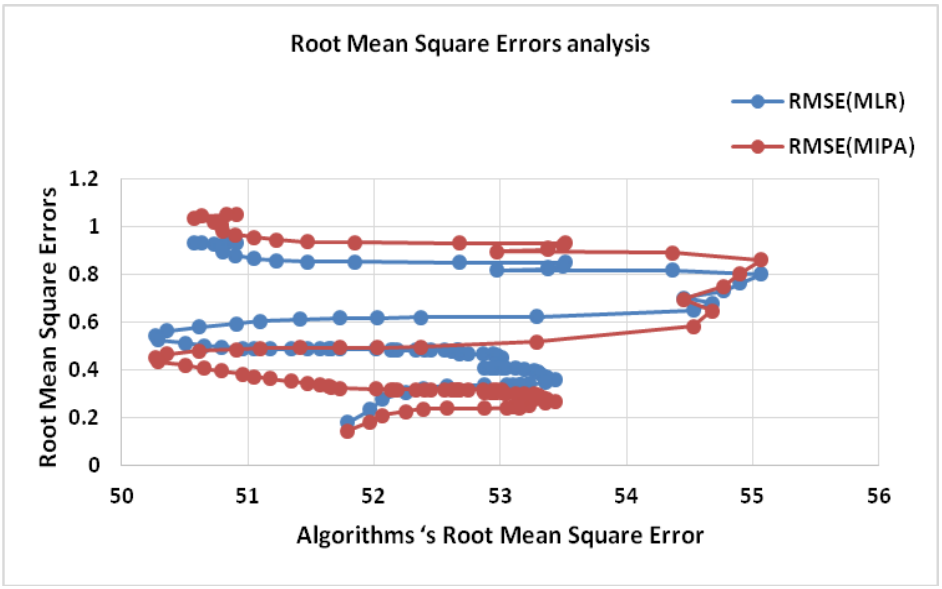


Figure 4. Analysis of Root Mean Square Errors (RMSE)

5.2. Analysis of Residuals

In Figure 5, the red line represents the plotting of residuals 0.99597124, 1.17010151, 1.48437316, respectively which are generated by the existing algorithm MLR, the blue line represents the plotting of residuals 0.421075736, 0.008795312, 0.553546256, respectively which are generated by MIPA. These values are corresponding to humidity. By analyzing Figure 5 it can be easily observed that the improved algorithm gives low error rates during the prediction of humidity based on temperature, wind speed, etc. independent variables.

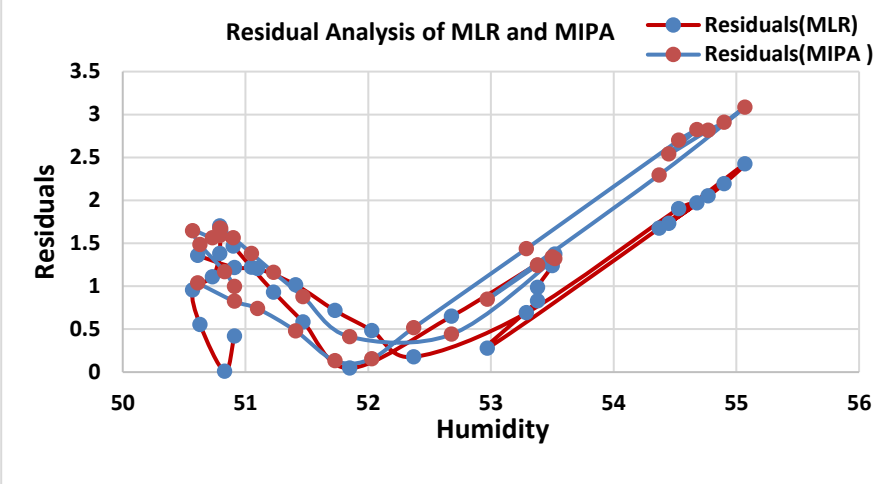


Figure 5. Analysis of residuals during prediction

5.3. Analysis of Mean Absolute Error (MAE)

MAE takes the absolute difference between the values that are actual and predicted and finds the average. MAE is crucial to identify the absolute value because it doesn't allow for any form of error value cancellation. For example, the average value of 0 if the average of 1 and -1 is considered because 1 and -1 will cancel out each other. In Figure 6 MAE for algorithms MLR and MIPA are compared, MIPA gives a better result. MIPA Prediction algorithm gives a better selection of important predictors or independent variables out of many independent variables in data sets. It reduces the cost of maintenance and collection of the data sets.

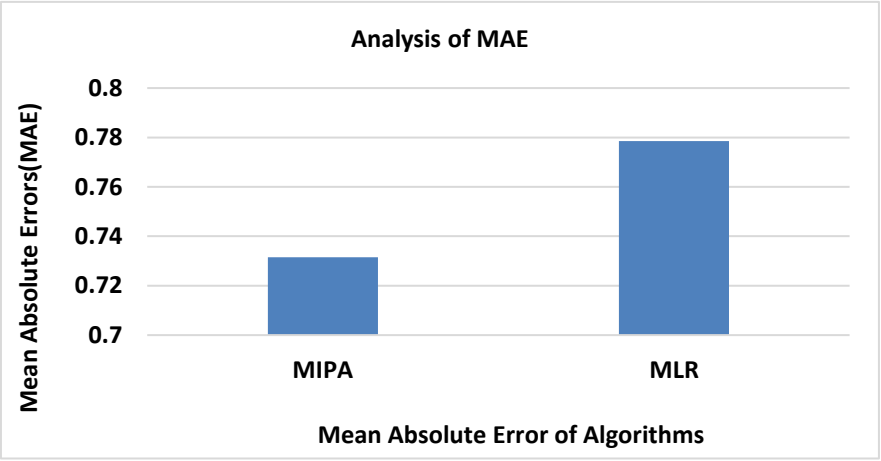


Figure 6. Analysis of Mean Absolute Error (MAE)

5.4. Analysis of Mean Absolute Percentage Error (MAPE)

MAPE is often referred to as the mean absolute percentage deviation (MAPD), a measure of the prediction accuracy of a statistical forecasting system. In Figure 7 MAPE for algorithms MLR and MIPA are compared, it gives a better result.

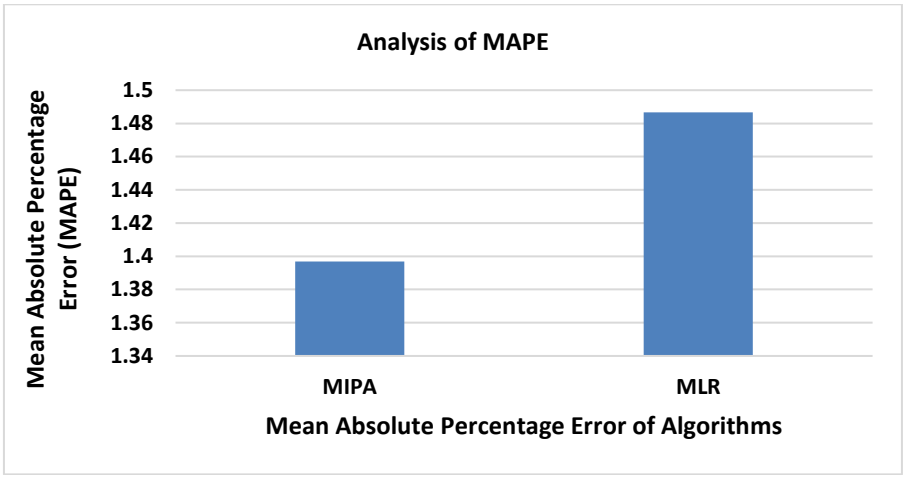


Figure 7. Analysis of Mean Absolute Percentage Error (MAPE)

5.5. Low cost in execution and maintenance of data sets

In the MIPA algorithm, only relevant variables are considered for data analysis and prediction process. The irrelevant variables get eliminate after the process of selection of variables or preprocessing, by eliminating irrelevant variables maintenance costs of data get reduce and it takes less time in the execution. Initially, in the data set four independent variables X_1, X_2, X_3 and X_4 are used, after the preprocessing of data set X_1, X_2, X_3 selected as relevant independent variables. The irrelevant variable X_4 gets eliminated. For further prediction process, we do not need to maintain the X_4 as the independent variable, so it takes less time in the prediction process.

6. Conclusion and Future Scope

In this paper, a MIPA algorithm is based on actual mean values. The analysis of the algorithms is done based on the parameters such as RMSE and residuals, MAE, MAPE. The accuracy of the prediction algorithm is measured with low RMSE and residuals. In this algorithm deviation for actual means is estimated for each relevant independent variable, using this estimated value, the prediction algorithm is framed. Prediction algorithm for 'n' independent variables is framed and it predicts the "Humidity" based on the rest of the independent variables. In Figure 3 it can be easily analyzed that average residuals of MIPA are 5.11% less than MLR, 24.62% are less than FIPM and 16.46 % are less than FTPDS. In section 5 it has been observed that MIPA is better than MLR.

MIPA is the regression based algorithm that can also be used in medical areas for the prediction of diseases based on the symptoms of patients. Through analysis, it can be found that values of error rate are more reduced in the implemented regression algorithm rather than MLR. It reduces the cost of data maintenance and reduces the execution time. It can help in the forecasting of diseases, revenues of the company, production, and weather, and in other areas.

Author Contributions: Each author has participated and contributed adequately to Take open accountability for suitable portions of the content.

Funding: This research received no external funding.

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

References

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J.-M. (2021). Random forests for global sensitivity analysis: A selective review. Reliability Engineering & System Safety. 28 193 – 222.

Chai, D. J., Kim, E. H., Jin, L., Hwang, B., & Ryu, K. H. (2007). Prediction of Frequent Items to One Dimensional Stream Data. International Conference on Computational Science and its Applications(ICCSA). 353 – 360. IEEE.

Daniya, T., Geetha. M., & Cristin, B. (2020).Least Square Estimation of Parameters for Linear Regression. International Journal of Control and Automation, 13, 447 - 452.

Gauba, H., Kumar, P., Roy, P. P., Singh, P., Dogra, D. P., & Raman, B. (2017). Prediction of advertisement preference by fusing EEG response and sentiment analysis. Neural Networks, 92, 77–88 .

Ilayaraja M., & Meyyappan T. (2015). Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets. Procedia Computer Science,70, 586–592.

Kavitha S, Varuna S ., & Ramya R.(2016). A comparative analysis on linear regression and support vector regression, International Conference on Green Engineering and Technologies (IC-GET), (1-5).IEEE.

Khan, F., Kari, D., Karatepe, I. A., & Kozat, S. S. (2016). Universal Nonlinear Regression on High Dimensional Data Using Adaptive Hierarchical Trees. IEEE Transactions on Big Data, 2(2), 175–188.

Mukherjee, S., Ghosh, S., Ghosh, S., Kumar, P., & Roy, P. P. (2019). Predicting Video-frames Using Encoder-convlstm Combination. International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2027-2031). IEEE.

Mustapha, A., & Fadzil, F. (2015). A Regression Approach for Forecasting Vendor Revenue in Telecommunication Industries. International Journal of Engineering and Technology. 6, 2604-2608.

Ostertagova, E., Frankovsky, P., & Ostertag, O. (2016). Application of polynomial regression models for prediction of stress state in structural elements. Global Journal of Pure and Applied Mathematics. 12, 3187-3199.

Saptawati, G. A. P., & Nata, G. N. M. (2015). Knowledge discovery on drilling data to predict potential gold deposit. International Conference on Data and Software Engineering (ICoDSE), (143-147). IEEE.

Yang, X., Mao, S., Gao, H., Duan, Y., & Zou, Q. (2019). Novel Financial Capital Flow Forecast Framework Using Time Series Theory and Deep Learning: A Case Study Analysis of Yu'e Bao Transaction Data. IEEE Access, 7, 70662–70672.

Yıldırım, D.C., Toroslu, I.H. & Fiore, U. (2021). Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators. Financ Innov 7.1-36.

Zhao. F., & Li, Q. (2005). A plane regression-based sequence forecast algorithm for stream data. International Conference on Machine Learning and Cybernetics(ICMLC), (1559-1562). IEEE.