PREDICTIVE ANALYSIS FOR MARKET SALES USING REGRESSION TECHNIQUE

[1]Saniya fatima, [2]Dr. V.K. Senthil Ragavan
[1]PG Scholar, M.Tech, Dept of CSE, Shadan Women's College of Engineering and Technology, Hyderabad, TS.
saniya98fatima@gmail.com
[2]Professor, Dept of CSE, Shadan Women's College of Engineering and Technology, Hyderabad, TS.
vksenrag@yahoo.com

ABSTRACT

Currently, supermarket run-centres, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting the sales of a business such as Big-Mart, and it was discovered that the model outperforms existing models.

1. INTRODUCTION

Everyday competitiveness between various shopping entrées as and as huge marts is becoming higher intense, violent just because of the quick development of global malls also online shopping. Each market seeks to offer personalized and limited-time deals to attract many clients relying on period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services. The current machine learning algorithm is very advanced and provides methods for predicting or forecasting sales any kind of organization, extremely beneficial to overcome low – priced used for prediction. Always better prediction is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful.

Moving Average, (ARMA) Auto-Regressive Moving Average, have been utilized to develop a few deals forecast standards. Be that as it may, deals anticipating is a refined issue and is influenced by both outer and inside factors, and there are two significant detriments to the measurable technique as set out in A. S. Weigend et A mixture occasional quantum relapse approach and (ARIMA) Auto-Regressive Integrated Moving Average way to deal with every day food deals anticipating were recommend by N. S. Arunraj and furthermore found that the exhibition of the individual model was moderately lower than that of the crossover model

What is Machine Learning?

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to predict an output which can be used to make actionable insights.

The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers.

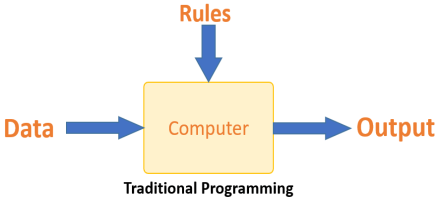
A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

Machine Learning vs. Traditional Programming

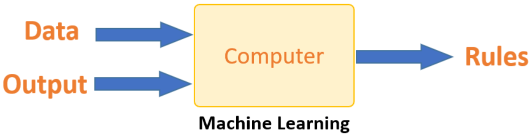
Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.



Traditional Programming

Machine learning is supposed to overcome this issue. The machine learns how the input and output data are correlated and it writes a rule. The programmers do not need to write new rules each time there is new data. The algorithms adapt in response to new data and experiences to improve efficacy over time.



Machine Learning

2. RELATED WORK

Suresh K and Praveen O, 2020. Information mining strategies are utilized broadly in commercial areas for separating data from the database. Information mining comprises of applying Utility Pattern Mining indicates about time utilizing strategies with considering of item sets. Utility Pattern Mining (UPM) is reasonable for significant data which assesses effectively in pattern identification. In this research paper, Hierarchical High Average Utility Pattern Mining (HAUPM) is proposed for e-commerce and retail industries. Unbounded stream data may generate constant outcomes which are needed to be updated based on time factor. Hierarchical High Average Utility Pattern Mining (HAUPM) used to perform operations on unbounded stream information on the database. Upon which state-of-the-art algorithm is performed on information which has a higher impact than more recent information. These data sets give profitable outcomes for retail industry based by making customers buy products which are trending in the market. H-HAUPM is been choose over other techniques is to obtain items that have a high impact based on accuracy in generating itemsets, not consuming more space for usage, scalability and maintaining consistency.

Suma, V., and Shavige Malleshwara Hills, 2020. There has been an increasing demand in the e-commerce market for refurbished products across India during the last decade. Despite these demands, there has been very little research done in this domain. The real-world business environment, market factors and varying customer behavior of the online market are often ignored in the conventional statistical models evaluated by existing research work. In this paper, we do an extensive analysis of the Indian e-commerce market using data-mining approach for prediction of demand of refurbished electronics. The impact of the real-world factors on the demand and the variables are also analyzed. Real-world datasets from three random e-commerce websites are considered for analysis. Data accumulation, processing and validation is carried out by means of efficient algorithms. Based on the results of this analysis, it is evident that highly accurate prediction can be made with the proposed approach despite the impacts of varying customer behavior and market factors. The results of analysis are represented graphically and can be used for further analysis of the market and launch of new products.

Shobha Rani, N., Kavyashree, S., & Harshitha, R, 2020. A lot of attention has emerged regarding the aspects of text detection and identification as OCR has generated a lot of prominence over the years. There has been a number of experiments conducted in this field to make the results more and more accurate. Most of the experiments carried out have paid attention to only a few attributes and not a lot of trails have been done for unusual scenarios, like a lot of techniques produces accurate results only for horizontal textual orientation. So there should different techniques for analyzing such images which have a complex background, different font styles, colors, textual orientations. Text detection on images containing texts of different orientations, different font types, and images with complex backgrounds is taken for the proposed work. There are mainly 3 steps in the algorithm proposed, the Canny edge detection approach for gradient filtering is applied in the first stage to detect the skeletal structure of various objects in the image. In the next stage textual threshold-based object filtering is carried out using the convolution technique with a heuristic thresholding model. The textual object filtering after convolution is subject to the last stage called post enhancement technique. In this stage, partial non-textual objects being filtered out are employed for removal based on geometrical properties of gradients of images, thus retaining only the textual objects. Finally, the textual object filtered gradient image is considered as a mask image for mapping it to the original image for text detection. Experimentations are conducted on Google Street View Datasets for which a subjective evaluation procedure is adapted to validate the results resulting in promising outcomes for more than 50% of images.

Wang, Haoxiang, 2019. Combination of Green supply chain management, green product deletion decision and green cradle-to-cradle performance evaluation with Adaptive-Neuro-Fuzzy Inference System (ANFIS) to create a green system. Several factors like design process, client specification, computational intelligence and soft computing are analyzed and emphasis is given on solving problems of real domain. In this paper, the consumer electronics and smart systems that produce nonlinear outputs are considered. ANFIS is used for handling these nonlinear outputs and offer sustainable development and management. This system offers decision making considering multiple objectives and optimizing multiple outputs. The system also provides efficient control performance and faster data transfer.

Giuseppe Nunnari, Valeria Nunnari, 2017. This paper presents a case study concerning the forecasting of monthly retail time series recorded by the US Census Bureau from 1992 to 2016. The modeling problem is tackled in two steps. First, original time series are de-trended by using a moving windows averaging approach. Subsequently, the residual time series are modeled by Non-linear Auto-Regressive (NAR) models, by using both Neuro-Fuzzy and Feed-Forward Neural Networks approaches. The goodness of the forecasting models, is objectively assessed by calculating the bias, the MAE and the RMSE errors. Finally, the model skill index is calculated considering the traditional persistent model as reference. Results show that there is a convenience in using the proposed approaches, compared to the reference one.

3. METHODOLOGIES

Always better prediction is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful.

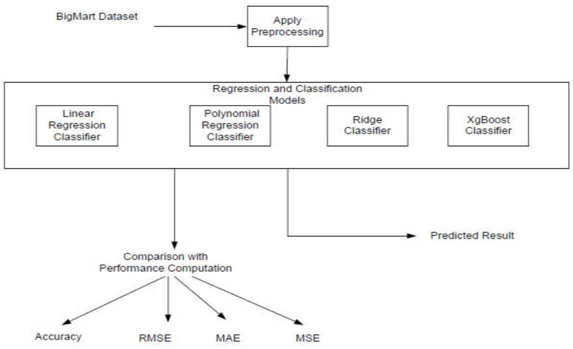

BigMart Dataset
Apply Preprocessing
Regression and Classification Models
Linear Regression Classifier
Polynomial Regression Classifier
Ridge Classifier
XgBoost Classifier
Predicted Result
Comparison with Performance Computation
Accuracy
RMSE
MAE
MSE

Fig1: Shows the proposed Architecture Diagram

- ➢ Data Collection
- ➢ Dataset
- ➢ Data Preparation
- ➢ Model Selection
- ➢ Analyse and Prediction
- ➢ Accuracy on test set
- ➢ Saving the Trained Model

Data Collection:
This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Dataset:
The dataset consists of 8523individual data. There are 12 columns in the dataset, which are described below.
1.ItemIdentifier ---- Unique product ID
2.ItemWeight ---- Weight of product
3.ItemFatContent ---- Whether the product is low fat or not
4.ItemVisibility ---- The % of the total display area of all products in a store allocated to the particular product
5.ItemType --- The category to which the product belongs
6.ItemMRP ---- Maximum Retail Price (list price) of the product
7.OutletIdentifier ---- Unique store ID
8.OutletEstablishmentYear ---- The year in which the store was established
9.OutletSize ---- The size of the store in terms of ground area covered
10.OutletLocationType ---- The type of city in which the store is located
11.Outlet Type ---- Whether the outlet is just a grocery store or some sort of supermarket
12.ItemOutletSales ---- sales of the product in t particular store. This is the outcome variable to be predicted.

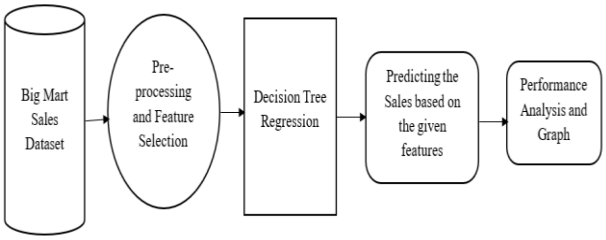
Data Preparation:
Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, and data type conversions, etc.)
Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
Split into training and evaluation sets

Model Selection:
We used decision tree regression machine learning algorithm, we got a accuracy of 95.7% on test set so we implemented this algorithm.

Decision Tree Regression
Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. The branches/edges represent the result of the node and the nodes have either:

Conditions [Decision Nodes]
Result [End Nodes]
The branches/edges represent the truth/falsity of the statement and take makes a decision based on that in the example below which shows a decision tree that evaluates the smallest of three numbers:

Decision Tree Regression: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.


Big Mart Sales Dataset
Pre-processing and Feature Selection
Decision Tree Regression
Predicting the Sales based on the given features
Performance Analysis and Graph

4. PROPOSED ALGORITHM EXPERIMENTAL RESULT ANALYSIS

A. Linear Regression:

- Build a fragmented plot.1) a linear or non-linear pattern of data and 2) a variance (outliers). Consider a transformation if the marking isn't linear. If this is the case, outsiders, it can suggest only eliminating them if there is a non-statistical justification.
- Link the data to the least squares line and confirm the model assumptions using the residual plot (for the constant standard deviation assumption) and the normal probability plot (for the normal probability assumption) A transformation might be necessary if the assumptions made do not appear to be met.
- If required, convert the data to the least square using the transformed data, construct a regression line.
- If a change has been completed, return to the previous process 1. If not, continue to phase 5.
- When a "good-fit" classic is defined, write the least-square regression line equation. Consist of normal estimation, estimation, and Rsquared errors.

Linear regression formulas look like this:

$$Y=o_1x_1+o_2x_2+\dots\dots\dots o_nx_n$$

R-Square: Defines the difference in X (depending variable) explains the total variance in Y (dependent variable) (independent variable). This can be expressed mathematically as

$$R-Square=1-\frac{\sum(Y_{actual}-Y_{predicted})\hat{}2}{\sum(Y_{actual}-Y_{mean})\hat{}2}$$

B. Polynomial Regression Algorithm

Polynomial Regression is a relapse calculation that modules the relationship here among dependent(y) and the autonomous variable(x) in light of the fact that as most extreme limit polynomial. The condition for polynomial relapse is given beneath: y= b0+b1x_1+ $b_2x_1{}^2$+ $b_2x_1{}^3$+...... $b_nx_1{}^n$

- It is regularly alluded to as the exceptional instance of various straight relapse in ML. Since we apply some polynomial terms to the numerous straight relapse condition to change it to polynomial relapse adjustment to improve accuracy.
- The informational collection utilized for preparing in polynomial relapse is of a non-straight nature.
- It uses a linear regression model to fit complex and non-linear functions and datasets.

C. Ridge Regression

Ridge regression is a model tuning tool used to evaluate any data that suffers from multicollinearity. This method performs the L2 regularization procedure. When multicollinearity issues arise, the least squares are unbiased and the variances are high, resulting in the expected values being far removed from the actual values. The cost function for ridge regression:

Min(||Y – X(theta)||^2 + λ||theta||^2)

D. XGBoost Regression

“Extreme Gradient Boosting” is same but much more effective to the gradient boosting system. It has both a linear model solver and a tree algorithm. Which permits “xgboost” in any event multiple times quicker than current slope boosting executions. It underpins various target capacities, including relapse, order and rating. As "xgboost" is extremely high in prescient force however generally delayed with organization, it is appropriate for some rivalries. It likewise has extra usefulness for cross-approval and finding significant factors.

5. RESULT

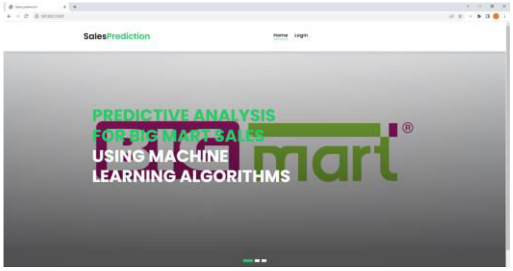

SalesPrediction
Home
Login
PREDICTIVE ANALYSIS FOR BIG MART SALES USING MACHINE LEARNING ALGORITHMS
bigmart

Fig 6. Home Page

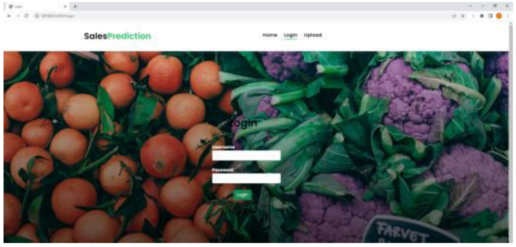


Fig 7. Login Page

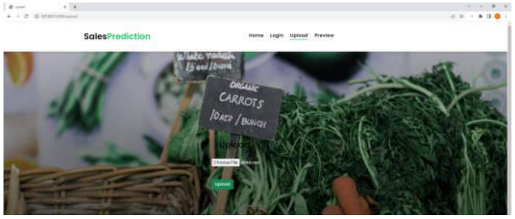
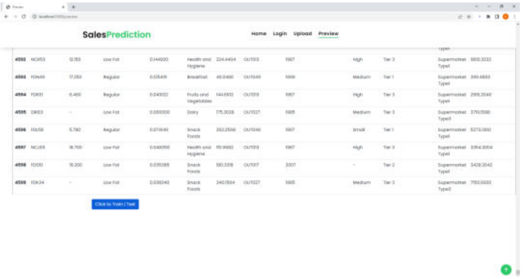
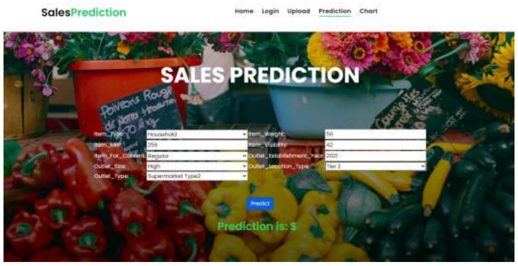


Fig 8. Upload Dataset Page



4592	NCR53	12.150	Low Fat	0.144920	Health and Hygiene	224.4404	OUT013	1987	High	Tier 3	Supermarket Type1	1800.3232
4593	FDN09	17.250	Regular	0.025419	Breakfast	45.0480	OUT049	1999	Medium	Tier 1	Supermarket Type1	399.4800
4594	FDR31	6.460	Regular	0.049122	Fruits and Vegetables	144.6802	OUT013	1987	High	Tier 3	Supermarket Type1	2958.2040
4595	DRI03	-	Low Fat	0.000000	Dairy	175.3028	OUT027	1985	Medium	Tier 3	Supermarket Type3	3710.1588
4596	FDL58	5.780	Regular	0.274849	Snack Foods	262.2568	OUT046	1997	Small	Tier 1	Supermarket Type1	5273.1300
4597	NCJ19	18.700	Low Fat	0.046056	Health and Hygiene	151.9920	OUT013	1987	High	Tier 3	Supermarket Type1	3354.3004
4598	FDS10	19.200	Low Fat	0.035288	Snack Foods	180.3318	OUT017	2007	-	Tier 2	Supermarket Type1	3428.2042
4599	FDK34	-	Low Fat	0.030340	Snack Foods	240.1584	OUT027	1985	Medium	Tier 3	Supermarket Type3	7150.6020

Fig 9. Train Page



Item_Type: Household
Item_Weight: 56
Item_MRP: 256
Item_Visibility: 42
Item_Fat_Content: Regular
Outlet_Establishment_Year: 2021
Outlet_Size: High
Outlet_Location_Type: Tier 2
Outlet_Type: Supermarket Type2

Prediction is: $

Fig 10. Prediction Page

6. CONCLUSION AND FUTURE ENHANCEMENT

In this work, the effectiveness of various algorithms on the data on revenue and review of, best performance-algorithm, here propose a software to using regression approach for predicting the sales cantered on sales data from the past the accuracy of linear regression prediction can be enhanced with this method, polynomial regression, Ridge regression, and Xgboost regression can be determined. So, we can conclude ridge and Xgboost regression gives the better prediction with respect to Accuracy, MAE and RMSE than the Linear and polynomial regression approaches. In future, the forecasting sales and building a sales plan can help to avoid unforeseen cash flow and manage production, staff and financing needs more effectively. In future work we can also consider with the ARIMA model which shows the time series graph.

Future Work: In future, the forecasting sales and building a sales plan can help to avoid unforeseen cash flow and manage production, staff and financing needs more effectively. In future work we can also consider with the ARIMA model which shows the time series graph.

REFERENCES

[1] Ching Wu Chu and Guoqiang Peter Zhang, “A comparative study of linear and nonlinear models for aggregate retails sales forecasting”, Int. Journal Production Economics, vol. 86, pp. 217- 231, 2003.

[2] Wang, Haoxiang. "Sustainable development and management in consumer electronics using soft computation." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 56.- 2. Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of D.

[3] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110.

[4] Giuseppe Nunnari, Valeria Nunnari, “Forecasting Monthly Sales Retail Time Series: A Case Study”, Proc. of IEEE Conf. on Business Informatics (CBI), July 2017. [

5] https://halobi.com/blog/sales-forecasting-five-uses/. [Accessed: Oct. 3, 2018].

[6] Zone-Ching Lin, Wen-Jang Wu, “Multiple Linear Regression Analysis of the Overlay Accuracy Model Zone”, IEEE Trans. on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 – 237, May 1999.

[7] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, “Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis”, Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 – 23, 2012.

[8] C. Saunders, A. Gammerman and V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables", Proc. of Int. Conf. on Machine Learning, pp. 515 – 521, July 1998.IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 56, NO. 7, JULY 2010 3561.

[9] "Robust Regression and Lasso". Huan Xu, Constantine Caramanis, Member, IEEE, and ShieMannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration." An improved Adaboost algorithm based on uncertain functions". Shu Xinqing School of Automation Wuhan University of Technology. Wuhan, China Wang Pan School of the Automation Wuhan University of Technology Wuhan, China.

[10] Xinqing Shu, Pan Wang, "An Improved Adaboost Algorithm based on Uncertain Functions", Proc. of Int. Conf. on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration, Dec. 2015.

[11] A. S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994.

[12] N. S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, Int. J. Production Economics 170 (2015) 321-335P.

[13] D. Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135.

[14] X. Yua, Z. Qi, Y. Zhao, Support Vector Regression for Newspaper/Magazine Sales Forecasting, Procedia Computer Science 17 (2013) 1055–1062.

[15] E. Hadavandi, H. Shavandi, A. Ghanbari, An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: a Case study of the printed circuit board, Expert Systems with Applications 38 (2011) 9392–9399.

[16] P. A. Castillo, A. Mora, H. Faris, J.J. Merelo, P. GarciaSanchez, A.J. Fernandez-Ares, P. De las Cuevas, M.I. Garcia-Arenas, applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment, Knowledge-Based Systems 115 (2017) 133-151.

[17] R. Majhi, G. Panda and G. Sahoo, "Development and performance evaluation of FLANN based model for forecasting of stock markets". Expert Systems with Applications, vol. 36, issue 3, part 2, pp. 6800-6808, April 2009.

[18] Pei Chann Chang and Yen-Wen Wang, "Fuzzy Delphi and back propagation model for sales forecasting in PCB industry", Expert systems with applications, vol. 30, pp. 715-726, 2006.

[19] R. J. Kuo, Tung Lai HU and Zhen Yao Chen "application of radial basis function neural networks for sales forecasting", Proc. of Int. Asian Conference on Informatics in control, automation, and robotics, pp. 325- 328, 2009.

[20] R. Majhi, G. Panda, G. Sahoo, and A. Panda, "On the development of Improved Adaptive Models for Efficient Prediction of Stock Indices using Clonal-PSO (CPSO) and PSO Techniques", International Journal of Business Forecasting and Market Intelligence, vol. 1, no. 1, pp.50-67, 2008.

[21] Suresh K and Praveen O, "Extracting of Patterns Using Mining Methods Over Damped Window," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 235-241, DOI: 10.1109/ICIRCA48905.2020.9182893.

AUTHOR PROFILE's

Ms. SANIYA FATIMA, Completed her B.Tech (CSE) from Methodist College Of Engineering And Technology, Abid's, Hyderabad, TS, India. Presently, she is pursuing her Masters in Computer Science &amp; Engineering from Shadan Women's College of Engineering & Technology, Khairtabad, Hyderabad, TS, India.

Dr. V K SENTHIL RAGAVAN is working as Professor of CSE in Shadan Women's College of Engineering and Technology, Hyderabad. He has obtained his Ph.D. in Computer Science and Engineering from Anna University, Chennai, Tamilnadu, India. Also, he was awarded with D.Sc. in Computer Science by Corllins University. He has 22 years of teaching experience and 1 year of industry experience. He has published 12 papers in National conferences and journals, 16 papers in International Conferences and Journals. He has published 4 patents and 2 books. He is a member of ISTE, SPIE, IETE and IEEE. He has organized various conferences, seminars, workshops and FDPs during his tenure. His area of interest includes Data Structures and Algorithms, Network Security, Artificial Intelligence and Image Processing.