Performance Improved Holt-Winter's (PIHW) Prediction Algorithm for Big Data Environment

B.Arputhamary[1], Dr. L.Arockiam[2]

[1] Mother Teresa Women's University, Kodaikanal, India.
[2] St. Joseph,s College, Tiruchirappalli, India.

Abstract

Prediction plays an important role everywhere particularly in business, technology and many others. It helps organizations to take timely decisions, to improve profits and to reduce lost sales. Recent years have witnessed an enormous development in the area of cloud computing and big data, which brings up challenges in decision making process. As the size of the dataset becomes extremely big, the process of extracting useful information by analysing these data has also become tedious. Today data are generated in an unprecedented manner, prediction plays major role in utilizing these data. Time Series based prediction models take great part in handling Big Data such as online sales data, weather data etc. In this paper a methodology for prediction is introduced and the model is evaluated by applying various time series models with time series data which is seasonal and non-stationary. From the analysis it is proved that Holt-Winter's model performs better in seasonal and non-stationary time series data. The Holt-Winters (HW) methods estimate three smoothing parameters, associated with level, trend and seasonal factors. The seasonal variation can be of either an additive or multiplicative form. Also in this paper, Performance Improved Holt-Winters (PIHW) prediction algorithm is proposed and the results demonstrate that a considerable reduction in forecast error (Mean Square Error) can be achieved in the proposed model compared to Holt-Winters (HW) model.

Keywords: Big Data, Holt-winter, Hadoop, MapReduce, Prediction

I. INTRODUCTION

Many household products are retailed by various organisations of the retail store network which are geographically located at different locations. Supply chain disorganisations will occur at different locations when the market potential will not evaluated by the retailers. Many times it is not easy for the retailers to understand the market condition at various geographical locations. The organization of retail store network has to understand the market conditions to intensify its goods to be bought and sold so that many number of customers get attracted in that direction. Business forecast helps retailers to visualize the big picture by forecasting the sales. The general ideas of forth coming years are needed if there is great change. These changes are achieved in the retail store's objective so that success is achieved more profitably. It also helps the customers to be happy by providing the products desired by them in desired time. When the customers are happy then the demand for that particular item will increase which will automatically increase the profit. The forecasting of sales will help the retailers to know about the demand of the product. Today data are generated tremendously and prediction plays important role in taking timely decisions. Time series analysis plays important role in forecasting the future with the data which has seasonal fluctuations. Time series analysis has different approaches for forecasting the future, exponential smoothing methods are very commonly used for forecasting demand. Exponential smoothing techniques are simple, fast and inexpensive [1]. The Holt-Winters (HW) methods can perform well with three smoothing parameters, associated with level, trend and seasonal factors. The seasonal variation can be of either an additive or multiplicative form. The multiplicative approach is widely used one which works better than the additive, but the multiplicative HW method may not be used if a data series contains some values equal to zero. In this paper, Performance Improved Holt-Winter's method is proposed and which treat the initial values for the level, trend and seasonal components as well as three smoothing constants (Alpha, Beta and Gamma) as decision variables. The proposed algorithm is compared with the

Holt-Winter's time series based prediction algorithm and the results shows that a considerable reduction in forecast error (Mean Square Error) is achieved.

II. Big Data Analytics

The term "Big Data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. Big Data sizes are constantly increasing from a few dozen terabytes to many petabytes of data in a single data set. In 2010, Apache Hadoop[5] defined big data as "Datasets which could not be captured, managed and processed by general computers within an acceptable scope. Gartner defined Big Data with 3 V's model: Volume, Velocity and Variety [5],[6]. Volume describes the generation and collection of massive amount of data as well as the data scale which becomes increasingly big. Velocity is the timeliness of big data and Variety means various types of data which includes semi structured and unstructured data such as audio, video, web page and text. Today data are generated in an unprecedented manner. These data are generated through many sources such as web logs, social networks (Blogs, Comments and Likes), transactional data sources and sensor data. The data obtained through various sources are heterogeneous in nature. Due to its nature, big data has generated a number of challenges in the decision making process. Today organizations are struggling in capturing, storing and analyzing these high volumes of data to increase the accuracy of decision making. Storing these voluminous data does not pose much problem but the effective utilization of these stored data is another challenge focused today. The challenges like scalability, unstructured data accessibility, real time analytics, fault tolerance and many more are handled by traditional approaches which have proved to be less efficient. Predictive analytics plays important role in handing such data (ie.) Big Data[7][11].

III. Forecasting Techniques

Forecasting is the process of extracting meaningful information from analyzing the historical data. There are many ways for forecasting the demand of various goods or services. Figure 1 illustrates how the various techniques can be classified[8][9].

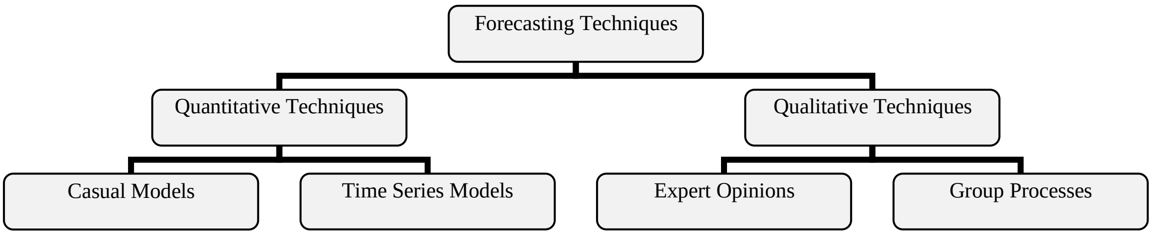


Fig.1. Classification of Forecasting Techniques

A. Quantitative Techniques

Quantitative techniques are statistical approaches which depends on the historical data. Casual models and Time series models are most important in quantitative approaches. Casual models are devoloped by using linear regression approach based on the method of least square. Time series demand is viewed as a sequence of observed values, which can be of the following form.

$$A_1,\ A_2,\cdots,A_n$$

These models attempt to identify patterns that have been present in the past and assume they will continue in the future. These models are often termed *extrapolation* models. A time series that has no trend is called *stationary*; if trend is present the time series is *nonstationary*.

B. Qualitative Techniques

Qualitative techniques depend on the experience that has not been captured in the form of hard data.

C. The Forecasting Process

There are three general steps in forecasting process. They are,

- Use past data to estimate the parameters of the model.

- Use estimated parameters to determine how well the time series model would have done in predicting past demand.
- Use estimated parameters to forecast demand for the future.

IV. Time series forecasting Techniques

This section first presents an overview on the widely used time series models for prediction, secondly the details of the methodology used for online sales forecasting is presented and finally, a recommendation to choose an appropriate prediction method according to seasonal conditions is given.

A. Autoregressive Integrated Moving Average (ARIMA) Models

In ARIMA models a non-stationary time series is made stationary by applying finite differencing of the data points. The mathematical formulation of the ARIMA(p,d,q) model using lag polynomials is given below[2][3]:

$$\phi(L)(1-L)^d y_t = \theta(L)\varepsilon_t, i.e.$$

$$\left[1-\sum_{i=1}^{p}\phi_i L^i\right](1-L)^d y_t = \left[1+\sum_{j=1}^{q}\theta_j L^j\right]\varepsilon_t \qquad (1)$$

Here, p, d and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively.

• The integer d controls the level of differencing. Generally d=1 is enough in most cases. When d=0, then it reduces to an ARMA(p,q) model.

• An ARIMA(p,0,0) is nothing but the AR(p) model and ARIMA(0,0,q) is the MA(q) model.

• ARIMA(0,1,0), i.e. $y_t = y_{t-1} + \varepsilon_t$ is a special one and known as the Random Walk model. It is widely used for non-stationary data, like economic and stock price series.

B. Artificial Neural Network(ANN) Model

Artificial Neural Network (ANN) has been suggested as an alternative technique to time series forecasting and it is gained immense popularity in last few years. ANN has been successfully applied in many different areas, especially for forecasting and classification process. During the past few years, a substantial amount of research works have been carried out towards the application of neural networks for time series modeling and forecasting. Similar to the work of the human brain, ANN_S try to recognize regularities and patterns in the input data, learn from experience and then provide generalized results based on their known previous knowledge.

Benefits of ANN models are,

- ANN_S are data driven and self-adaptive in nature
- ANN_S are inherently non linear
- Universal, functional approximation

C. Support Vector Machine(SVM) model

Support Vector Machine is developed by Vapnik and his co-workers at the AT & T Bell Laboratories in 1995. The initial aim of SVM was to solve pattern classification problems but afterwards they have been applied in many fields, such as function estimation, regression, signal processing and time series prediction problems. The remarkable characteristics of SVM are that it is not only designed for good classification but also intended for a better generalization of the training data. For this reason the SVM methodology has become one of the well known techniques, especially for time series forecasting problems in recent years.

In SVM, the solution to particular problem only depends upon a subset of the training data points, which are termed as support vectors. The solution obtained by applying SVM method is always unique and globally optimal. In SVM, the input points are mapped to a high dimensional feature space, with the help of some special functions, known as support vector kernals.

The two popular SVM models for time series forecasting are,

i. Least Square SVM(LS-SVM)

ii. Dynamic Least Square SVM

D. Holt Winter's Algorithm

The exponential smoothing formulae applied to a series with a trend and constant seasonal component using Holt Winter's additive technique are [2][3],

Additive Model:

$$a_t = \alpha(Y_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}) \quad (2)$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \quad (3)$$

$$s_t = \gamma(Y_t - a_t) + (1-\gamma)s_{t-p} \quad (4)$$

where α, β and γ are the smoothing parameters.

a_t is the smoothed level at time t.

b_t is the change in the trend at time t.

s_t is the seasonal smooth at time t.

p is the number of seasons per year.

The initial values required for Holt Winter's algorithm are,

$$a_p = \frac{1}{p}(y_1 + y_2 + ... + y_p) \quad (5)$$

$$b_p = \frac{1}{p}\left[\frac{y_{p+1} - y_1}{p} + \frac{y_{p+2} - y_2}{p} + ... + \frac{y_{p+p} - y_p}{p}\right] \quad (6)$$

$$s_1 = Y_1 - a_p \; s_2 = Y_2 - a_p \; ...., \; s_p = Y_p - a_p \quad (7)$$

$$y_{T+\tau} = a_T + b_T + s_T \quad (8)$$

Multiplicative Model:

$$a_t = \alpha\frac{Y_t}{s_{t-p}} + (1-\alpha)(a_{t-1} + b_{t-1}) \quad (9)$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \quad (10)$$

$$s_t = \gamma\frac{Y_t}{a_t} + (1-\gamma)s_{t-p} \quad (11)$$

The initial values for multiplicative model are:

$$a_p = \frac{1}{p}(y_1 + y_2 + ... + y_p) \quad (12)$$

$$b_p = \frac{1}{p}\left[\frac{y_{p+1} - y_1}{p} + \frac{y_{p+2} - y_2}{p} + ... + \frac{y_{p+p} - y_p}{p}\right] \quad (13)$$

$$s_1 = \frac{Y_1}{a_p}, \; s_2 = \frac{Y_2}{a_p}, \; ...., \; s_p = \frac{Y_p}{a_p} \quad (14)$$

The Holt-Winters forecasts are then calculated using the latest estimates from the appropriate exponential smoothes that have been applied to the series. This experiment is conducted in R with the airline passenger dataset that represents the monthly total number of international airline passengers (in thousands) from January 1949 to December 1960. It is a famous time series and is used by many analysts including Box and Jenkins. The important characteristic of this series is that it follows a multiplicative seasonal pattern with an upward trend. The airline passenger series has total 144 observations, out of which we have used the first 12 months for training and the next 12 months for testing. Due to the presence of strong seasonal variations, the airline data is non-stationary. From the results it is concluded that Holt-Winter's model is an appropriate model for handling time series data which has more seasonal fluctuations. Therefore in further research contributions, the Holt-Winter's model is considered.

Table 1 Analysis of Time series models for non-stationary seasonal data

Methods	MSE	RMSE	MAPE

Holt-Winter
(Alpha = 0.266,
Beta = 0.056,
Gamma = 0.5)
	176.8853	13.299823	2.336608%

ARIMA	189.333893	13.759865	2.244234%
SVM $\begin{pmatrix} \sigma = 1.512\times 10^{7} \\ C = 1.277\times 10^{10} \\ n = 35,\ N = 97 \end{pmatrix}$	186.0153	13.6387	2.66221%
ANN (1, 12 13; 2)	2532.23	50.321353	8.454268%

From the analysis of time series prediction techniques, it concluded that Holt Winters prediction model is most appropriate for handling data with more seasonal fluctuations.

V. Prediction Methodology

Forecasting time series data is an integral component for management, planning and decision making. Following the Big Data trend, large amounts of time series data are available from many heterogeneous data sources in more and more applications domains. The highly dynamic and often fluctuating character of these domains in combination with the logistic problems of collecting such data from a variety of sources imposes new challenges to forecasting. Traditional approaches heavily rely on extensive and complete historical data to build time series models and are thus no longer applicable if time series are short or, even more important, intermittent. In addition, large numbers of time series have to be forecasted on different aggregation levels with preferably low latency, while forecast accuracy should remain high. This is almost impossible, when keeping the traditional focus on creating one forecast model for each individual time series. Therefore, it is necessary to propose a new technique to perform forecasting with time series Big Data.

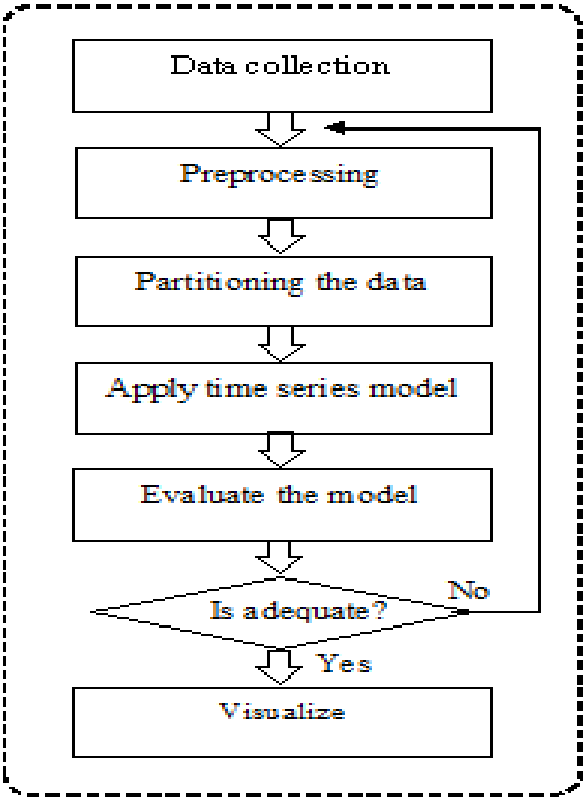


Fig. 2 Block diagram of a time series prediction model

In the proposed method, it is assumed that the numerical samples and their time stamps are stored on HDFS before invoking the input data. Data partitioning and pre-processing are handled by the map function. The partitioned data are then shuffled, sorted and aggregated with the same unique key. If all of data from map function is available at the reducer, the prediction is performed and its performance (in terms of prediction error) is measured. The final output is created by combining the results of all prediction. The following diagram illustrates the detailed steps of proposed algorithm in both of map and reduces stages [12].

VI. Holt-Winter's (HW) and Performance Improved Holt-Winter's (PIHW) Algorithm

Time Series forecasting assumes that a time series is a combination of a pattern and some random error. The goal is to separate the pattern from the error by understanding the pattern's trend, its long term increase or decrease (ie.) level and its seasonality, the change caused by seasonal factors such as fluctuations in use and demand. Several methods of time series forecasting are available such as Moving Average Method, Linear Regression with Time, Exponential Smoothing[5][[6]. This paper concentrates on the Holt Winter's Exponential Smoothing techniques as time series that exhibit seasonality. The Holt Winter model uses a modified form of exponential smoothing. It applies three exponential formulae to the series. Firstly, the level/mean is smoothed to give a local average value for the series. Next, the trend and finally seasonal sub-series is smoothed separately to give a seasonal estimate for each of the seasons. The exponential smoothing formulae applied to a series with a trend and constant seasonal component using Holt Winter's additive technique are given in Table 2. Holt-Winter's algorithm is a simple and easy to use approach and can be used in Big Data environment to perform prediction and to overcome the problem of data underutilization[7][8]. In this paper, Holt-Winter's algorithm is analyzed and a Performance Improved Holt-Winter's (PIHW) algorithm is proposed which is shown in Table 2.

The Holt Winter (HW) and the Performance Improved Holt Winter (PIHW) models are compared and the difference is observed. In PIHW, level, seasonality and trend are calculated and the smoothing parameters such as α , β and γ are doubled in order to improve the forecasting accuracy. And the value of Y_t is subtracted with 1 to stabilize the smoothed level and trend. The proposed model is compared with Holt-Winter's algorithm by using online eBay dataset and the proposed model shows some significant improvements in forecasting accuracy at additive level. The following section describes the comparison between the HW and PIHW and the results show that the proposed PIHW performs better than the Holt-Winter's prediction algorithm. Table 3 describes the Holt-Winter's forecasting results at additive level and Table 4 estimates forecasting of PIHW. Y_t is actual data (sales data from 2007 to 2014), a_t calculates level, b_t calculates trend and S_t is seasonality. F_t measures forecasting value and E_t represents estimated error. The smoothing parameters alpha, beta and gamma must lies between 0 and 1, they are given below. Table 3 and Table 4 summarise the actual and forecasted value of sales data set using additive model.

The forecasting error of time series models can be measured by using Mean Square Error (MSE) method which is given in equation (6.1).

$$MSE\left(Y,\hat{Y}\right)=\sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2} \qquad (15)$$

Table 2 Holt-Winter's(HW) & Performance Improved Holt-Winter's(PIHW)

Holt-Winter's(HW) Algorithm	**Performance Improved Holt-Winters(PIHW) Algorithm**
Additive Model: $a_t = \alpha(Y_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1})$ $b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$ $s_t = \gamma(Y_t - a_t) + (1-\gamma)s_{t-p}$ **where α , β and γ are the smoothing parameters.** a_t **is the smoothed level at time t.** b_t **is the change in the trend at time t.** s_t **is the seasonal smooth at time t.** p **is the number of seasons per year.** **The initial values required for Holt Winter's algorithm are,** $a_p = \frac{1}{p}(y_1 + y_2 + ... + y_p)$ $b_p = \frac{1}{p}\left[\frac{y_{p+1}-y_1}{p} + \frac{y_{p+2}-y_2}{p} + ... + \frac{y_{p+p}-y_p}{p}\right]$ $s_1 = Y_1 - a_p \; s_2 = Y_2 - a_p \; ...., \; s_p = Y_p - a_p$ $y_{T+\tau} = a_T + b_T + s_T$	**Additive Model:** $a_t = 2\alpha(Y_t - s_{t-p} - 1) + (1-2\alpha)(a_{t-1} + b_{t-1})$ $b_t = 2\beta(a_t - a_{t-1}) + (1-2\beta)b_{t-1}$ $s_t = 2\gamma(Y_t - a_t - 1) + (1-2\gamma)s_{t-p}$ **where α , β and γ are the smoothing parameters.** a_t **is the smoothed level at time t.** b_t **is the change in the trend at time t.** s_t **is the seasonal smooth at time t.** p **is the number of seasons per year.** **The initial values required for PIHW (Additive Model) are,** $a_p = \frac{1}{p}(y_1 + y_2 + ... + y_p)$ $b_p = \frac{1}{p}\left[\frac{y_{p+1}-y_1}{p} + \frac{y_{p+2}-y_2}{p} + ... + \frac{y_{p+p}-y_p}{p}\right]$ $s_1 = Y_1 - a_p \; s_2 = Y_2 - a_p \; ...., \; s_p = Y_p - a_p$ $y_{T+\tau} = a_T + b_T + s_T$

Table 3. Holt-Winter's (HW) Additive Model

	Yt	a_t	b_t	S_t	F_t	Et	Et*Et
2007_APR	3547.29			-1835.75			
2007_MAY	3752.96			-1630.08			
2007_JUN	3714.74			-1668.3			
2007_JUL	4349.61			-1033.43			
2007_AUG	3566.34			-1816.7			
2007_SEP	5021.82			-361.222			
2007_OCT	6423.48			1040.438			
2007_NOV	7600.6			2217.558			
2007_DEC	19756.21	5383.04167	1644.673	14373.17			
2008_JAN	2499.81	6812.42698	1632.543	-4015.42	3309.483	-809.673	655569.9
2008_FEB	5198.24	8386.0428	1629.223	-3106.46	5419.858	-221.618	49114.61
2008_MAR	7255.14	9957.35538	1625.96	-2622.27	7472.934	-217.794	47434.17
--							
--							
2014_NOV	56720.38	40402.6873	582.3842	12854.7	47285.8	9434.58	89011303
2014_DEC	120038.4	49732.2426	1075.224	58231.15	87141.22	32897.17	1.08E+09

Table 4.Performance Improved Holt-Winter's (PIHW) Additive Model

	Yt	a_t	b_t	S_t	F_t	Et	Et*Et
2007_AUG	3566.34			-1816.7			
2007_SEP	5021.82			-361.222			
2007_OCT	6423.48			1040.438			
2007_NOV	7600.6			2217.558			
2007_DEC	19756.21	5383.041667	1644.673	14373.17			
2008_JAN	2499.81	6596.607769	1596.093	-4097.8	3309.483	-809.673	655569.9
2008_FEB	5198.24	8208.468915	1597.87	-3011.23	5167.589	30.65061	939.4602
2008_MAR	7255.14	9801.091666	1597.279	-2546.95	7264.007	-8.86734	78.62971
--							
--							
2014_NOV	56720.38	41042.32881	995.4012	15677.05	47445.34	9275.04	86026366
2014_DEC	120038.4	56140.50603	2584.579	63896.88	93517.88	26520.51	7.03E+08

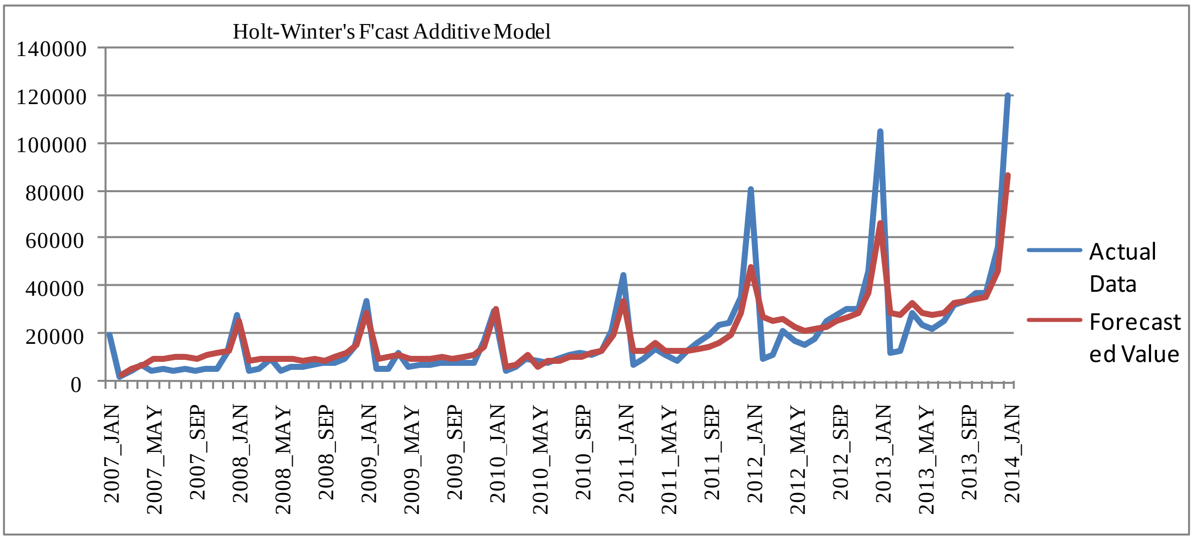


Fig 3. Holt-Winter’s forecasting

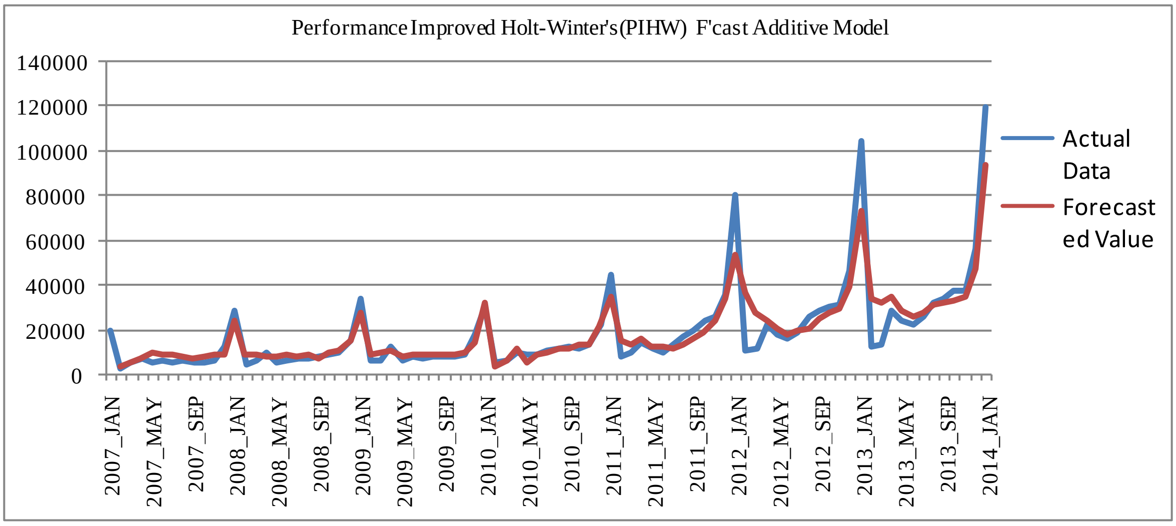
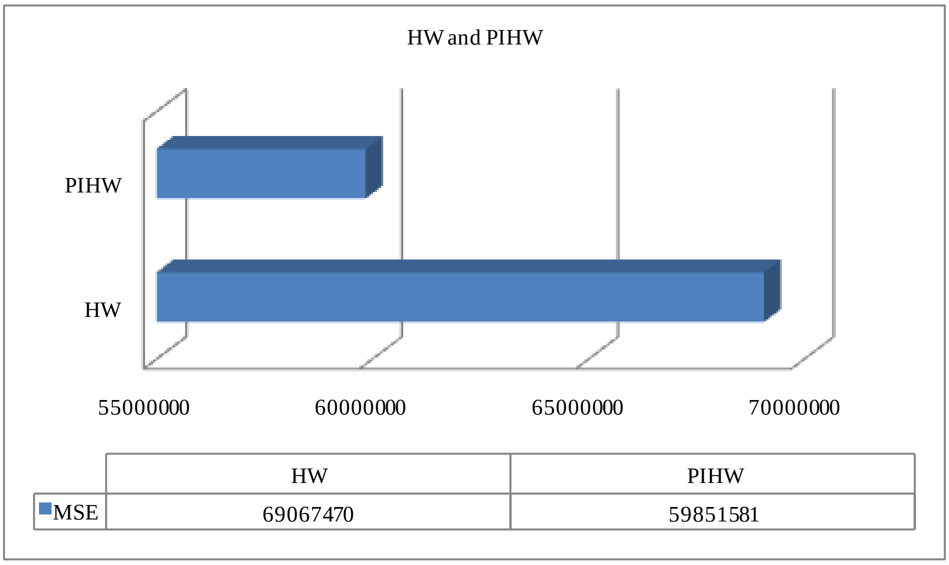


Fig 4. Performance Improved Holt-Winter’s (PIHW) forecasting



	HW	PIHW
■MSE	69067470	59851581

Fig 5. Performance Comparison Between HW and PIHW

The graph (Fig. 3.) shows the forecasting with respect to Holt-Winter's forecasting algorithm. And x-axis represents time series in months (from Jan-2007 to Dec-2014) and y-axis represents the actual sales range.

The above graph (Fig. 4.) shows the forecasting with respect to Performance Improved Holt-Winter's (PIHW) forecasting algorithm. And x-axis represents time series in months (from Jan-2007 to Dec-2014) and y-axis represents the actual sales range. The above figure 3 and 4 depicts the forecasting of HW and PIHW at additive level. The proposed Performance Improved Holt-Winter's(PIHW) performs better at additive level with respect to forecasting error. The following diagram depicts the performance comparison between HW and PIHW with respect to Mean Square Error (MSE).

In the above graph(Fig 5) x-axis represents the Mean Square Error(MSE) rate and the y-axix represents the forecasting error estimated in Holt-Winter's and Performance Improved Holt-Winter's(PIHW) algorithm. The result shows that PIHW perform better than the HW.

VII. CONCLUSION

Today the world is in Big Data era and people are trying to resolve the problem of handling Big Data in many aspects. Prediction is the one of the solutions to solve the problem of data underutilization in Big Data environment. In this paper, time series based forecasting algorithm Holt-Winter's is analysed and a Performance Improved Holt-Winter's (PIHW) algorithm is proposed and it performs better than the existing Holt-Winter's at additive level. In future the proposed algorithm can be verified at multiplicative and other models of Holt-Winter's algorithm.

REFERENCES

[1]. Anita S. Harsoor, Anushree Patil, " Forecast of sales of walmart store using big data applications", IJRET: International Journal of Research in Engineering and Technology, 2015.

[2]. B. Arputhamary, L.Arockiam, R.Thamarai Selvi, "Analysis of Prediction Techniques in Time Series for Big Data Using R", International Conference on Engineering Technology and Science (ICETS'15), 2015.

[3]. A. Banerjee , M. Marcellino and Masten, "Forecasting with Factor-augmented Error Correction Models", International Journal of Forecasting, In Press, 2013.

[4]. M. Banbura and M. Modugno, "Maximum Likelihood Estimation of Factor Models on Datasets with Arbitrary Pattern of Missing Data", Journal of Applied Econometrics, 2014.

[5]. D. Bernstein, "Big Data's Greatest Power: Predictive Analysis", 2013.

[6]. Dilpreet Singh and Chandan K Reddy, "A Survey on Platforms of Big Data Analytics", Journal of Big Data, Springer Open Access, 2014.

[7]. Min Chen, Shiwen Mao, Yunhao Liu, " Big Data: A Survey", Springer Science + Business Media, New York 2014, Springer, 2014.

[8]. Rashmi Ranjan Dhall, B.V.A.N.S.S. Prabhakar Rao, " Shrinking the Uncertainty In Online Sales Precdiction With Time Series Analysis", ICTACT journal on soft computing: special issue on distributed intelligent systems and applications, october 2014.

[9]. Rob J Hyndman with contributions from George Athanasopoulos, SlavaRazbash, Drew Schmidt, Zhenyu Zhou, Yousaf Khan, ChristophBergmeir, and Earo Wang, "Forecast: Forecasting Functions for Time Series and Linear Models", 2014.

[10]. Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", International Conference on Communication, Information and Computing Technology(ICCICT), 2012.

[11]. J. Shaw, "Why "big data" is a big deal?", "Harvard Business Review", 2014.

[12]. B. Arputhamary, L.Arockiam, "Parallel Prediction Model for Big Data Using MapReduce Programming Model", International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, N0. 82, 2015.