Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

Sales Forecasting Using Machine Learning Models

Dr.K. Deepa, G.Raghuram
Professor, Department of IT, Sri Ramakrishna Engineering College
StudentDepartment of IT, Sri Ramakrishna Engineering College
Email: deepa.senthil@srec.ac.in

Abstract*: Intelligent Decision Analytical System requires decision analysis and predictions. Most of the business organizations heavily depend on a knowledge base and demand prediction of sales trends. The accuracy in sales forecast provides a big impact in business. Data mining techniques are very effective tools in extracting hidden knowledge from an enormous dataset to enhance accuracy and efficiency of forecasting. The paper briefly analysed the concept of sales data and sales forecast using machine learning models. Based on a performance evaluation, a best suited predictive model is suggested for the sales trend forecast. The sales data contains the sales details for 3 years sales across 1,115 stores. It contains 9 attributes such as Store, Day of Week, Date, Sales, Customer, Open (Yes-1, No-0), Promo (Yes-1, No-0), State Holiday (Yes-1, No-0), School Holiday (Yes-1, No-0). The data is analysed using various time series algorithms such as Arima model, Benchmark method (Seasonal Naïve Bayes) and Exponential smoothing method. The analysis makes use of complete dataset and make predictions for the upcoming 2 years. The forecasted data of each algorithm are compared, and the final efficient result set is identified for the prediction of sales. The studies found that compared to Exponential smoothing and seasonal naïve model, Arima model (1,1,0)(0,1,0) as best fit model, which shows maximum accuracy of 74% in forecasting and future sales prediction*.

Keywords—Data Mining,Machine learning models, Arima mode, Seasonal naïve Bayes, Exponential smoothing.

I. INTRODUCTION

One of the major objectives of this research work is to find out the reliable sales trend prediction mechanism which is implemented by using data mining techniques to achieve the best possible revenue. Today's business handles huge repository of data. The volume of data is expected to grow further in an exponential manner. Any forecast can be termed as an indicator of what is likely to happen in a specified future time frame in a field. Therefore, the sales forecast indicates as to how much of a product is likely to be sold in a specified future period in a specified market at specified price. Forecasting is the process of estimation of quantity, type and quality of future work e.g. sales. Various machine learning algorithms for forecasting are follows

1. Arima Model
2. Exponential Smoothing Method
3. Naïve Bayes Method
4. Random Forest method
5. Linear regression method
6. Holt winter method etc.

II. LITERATURE SURVEY

"**Intelligent Sales Prediction Using Machine Learning Techniques"**

The fashion store dataset for three consecutive years of sales data is used. The sales prediction is done for upcoming 3 years from 2015 to 2017. Exploratory analysis stages involved in the data mining model include data understanding, preparation, modelling, evaluation and deployment. The forecast is composed of a smoothed averaged adjusted for a linear trend. Then the forecast is also adjusted for seasonality. Machine learning algorithms such as Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boost Tree (GBT) are used in prediction of future sales.

Based on the performance, Gradient Boost Algorithm is provides 98% overall accuracy and the second stands Decision Tree Algorithms with nearly 71% overall accuracy and followed by Generalized Linear Model with 64% accuracy. Finally, it can be compared based on the empirical evaluation of the three chosen algorithm the best fit for the model is Gradient Boosted Tree which provides the maximum accuracy of prediction across all the algorithms.

"MachineLearning Models for Sales Forecasting"

"Rosemann Store Sales" dataset is used to predict the future sales. The calculations were conducted in the Python environment using the main packages pandas, sklearn, numpy, keras, matplotlib, seaborn. Analysis is

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

done using Jupyter notebook. Regression algorithm captures the patternsin the whole set of stores or products.The analysis includes the attributes such as meansales value of historical data, state and school holiday flags, distance from store to competitor's store, store assortment type are considered in prediction. Various machine learning models such as Random Forest, Neural network, Lasso regularization, Arima model and ExtraTree model are used to analyze the data. The models in the first level (ExtraTree, Lasso, Neural Network) have non-zero coefficients for their results.

The solution is based on three level models. On the first level, many models were based on the XGBoost machine learning algorithm. In the second level, models from Python scikit-learn package, Extra tree model, linear model as well as Neural network model. The results from the second level were summed with weights on the third level.The use of regression approaches for sales forecasting can givebetter results compared to time series methods. ExtraTree method provides more stacking weights for regressors compared to other approaches.
"Walmart's Sales Data Analysis- A Big Data Analytics Perspective"

The Walmart has 45 stores in geographically diverse locations, each of the store having 99 departments. The datasetcontains the weekly sales and the factors affecting sales such as (Temperature, fuel price, unemployment rate, holiday) for each store locations for 3 years. Apache data science platforms, libraries, and tools are used in this work. Tools like Hadoop Distributed File Systems (HDFS), Hadoop MapReduce framework and Apache Spark along with Scala, Java and Python high-level programming environments are used to analyse and visualize the data. Machine learning library is employed with a simple regression model to predict future sales.
The results are predicted from data analysis and based on the predicted results the Retailers need to plan and evaluate according to the market driving factors which are, and not limited to, the temperature, fuel prices holidays, human resources, geographical location and many more. Effective and efficient supply chain, inventory, human resource management is needed to avoid losing competitive edge in the market, especially planning sales at different locations

"Forecast of Sales of Walmart Store Using Big Data Applications"
The forecasting process uses Walmart sales data. The different types of stores such as convenient store, department store, luxury store, super market, shopping malls etc. helps in determining the business models and strategy of operations. The process involves the stages such as determine dependent and independent variables, develop forecast procedures, select forecast analysis method, gather and analyze data, present assumptions about data, make and finalize forecast, evaluate results.

The strategy includes the collection of huge data of sales and then it is transferred on HDFS (Hadoop distributed file system) and map reduced is performed on the data sets. The Holt winters algorithm is used to predict the sales .The seasonality, trend and randomness is observed in the algorithm .The algorithm is used for train data sets and then the sales prediction.

The final results represents that the numerical representation of the forecasted sales and the accuracy of sales predicted is measured by 80% low confidence sales, 80% high confidence sales and 95% low confidence sales and 95% high confidence sales , error factor can be found between the predicted sales and the observed sales data, i.e. to find the error factor of month June in both the predicted sales and the observed sales data then the difference between predicted sales and the observed sales data is obtained, if the difference between them is very low or negligible and thus the sales prediction will be more accurate.

III. EXISTING SYSTEM

Sales forecasting is usually done by collecting the sales data of a shop of a time period and make predictions using various prediction techniques. There are many factors which affects the sales forecasting which includes direct and indirect competition, state and city holidays, population changes, sales promotions etc. The above factors create a great deviation in sales prediction in existing system which is not providing accurate results as expected. The confidence level has not taken for all algorithms. The holiday factors which is important in sales prediction is no considered. Thus, the sales varies on using different machine learning algorithms.

IV. PROPOSED SYSTEM

In proposed system, various analysis method is used in order to predict data from the sales record of daily sales of 1,115 stores. The project makes use of the complete dataset and make predictions for the upcoming 6 weeks. Various algorithms such as Linear Regression, Time series analysis, Benchmark method (Seasonal Naive method), Exponential smoothing method, Arima model method are used to forecast the data, so that predictions

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

can be compared across all the results, to determine the most accurate result. The forecasted data of each algorithm are compared and the final efficient resultset is identified for the prediction of sales.

Time series data is used by the algorithm, so that data for the upcoming 3 years can be predicted with less variation across time and data. A time series can be broken down to its components to systematically understand, analyze, model and forecast it. Variance and effect of seasonality is monitored, so that it should not increase over time, which will increase chance to predict most accurate results.

Merits:
Time series data is used for sales prediction, which provides better forecasted results. Sales is forecasted using multiple machine learning algorithms and compared the results of each algorithms and found the best fit model. Minimum and maximum confidence level of 80% and 95% is obtained while forecasting the data which improvise the accuracy.

4.1. ARIMA MODEL

ARIMA, short for 'Auto Regressive Integrated Moving Average' is a class of models that explains a time series based on its own past values, that is, its own lags and the lagged forecast errors. Thus the equation can be used to forecast future values. It is characterized by 3 terms p, q, d.

p - order of AR term (Auto Regressive)
q - order of MA term (Moving Average)
d - number of differencing required to make the time series stationary

In ARIMA model, the time series is differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes,

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_1 + \Phi_1 \epsilon_{t-1} + \Phi_2 \epsilon_{t-2} + .. + \Phi_q \epsilon_{t-q} \text{---(1)}$$

Predicted Yt = Constant + Linear combination Lags of Y (p lags) + Linear Combination of Lagged forecast errors (q lags)

4.2.SEASONAL NAÏVE METHOD

In naïve forecasts, the forecasts are set to be the value of the last observation. That is,

$$\hat{y}_{T+h|T} = y_T. \text{ -----(2)}$$

This method works well for many economic and financial time series. Since naïve forecast is optimal when data follow a random walk, these are also called random walk forecasts**.**

For high seasonal data, the seasonal naïve method set each forecast to be equal to the last observed value from the same season of the year (e.g., the same month of the previous year). The forecast for time T+h is written as

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}, \text{ --------- (3)}$$

where m= the seasonal period, and k is the integer part of $(h-1) / m(h-1) / m$ (i.e., the number of complete years in the forecast period prior to time T+h), h is the forecast horizon (i.e. time limit for which forecast needs to be prepared) and T is the last observed data. This looks more complicated than it really is. For example, with monthly data, the forecast for all upcoming February values is equal to the last observed February value. With quarterly data, the forecast of all upcoming Q2 values is equal to the last observed Q2 value (where Q2 means the second quarter). Similar rules apply for other months, quarters and other seasonal periods.

V. EXPONENTIALSMOOTHING METHOD

Exponential smoothing forecasting methods are similar in that a prediction is a weighted sum of past observations, but the model explicitly uses an exponentially decreasing weight for past observations. Collectively, the methods are sometimes referred to as ETS models, referring to the explicit modelling of Error, Trend and Seasonality. Single Exponential smoothing is a time series forecasting for univariate data without trends or seasonality. It use single parameter alpha(α) which is called as smoothing factor. Alpha value always lies between 0 and 1.

The simplest form of exponential smoothing is given by the formula:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha) y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots, \text{ ---------(4)}$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

where α is the *smoothing factor*, and $0 < \alpha < 1$. In other words, the smoothed statistic $y_{t+1|T}$, is a simple weighted average of the current observation y_t and the previous smoothed statistic y_{t-1}. If the αvalue nearly equals to 1, then the forecast uses most recent observation of the dataset, and if the value nearly equals to 0, the it uses historical data for forecasting.
The trends and seasonality factor is removed by dampening the time series data. The trends and seasonality can be linear or exponential based on the time series data.

5.1.WORKING PROCESS

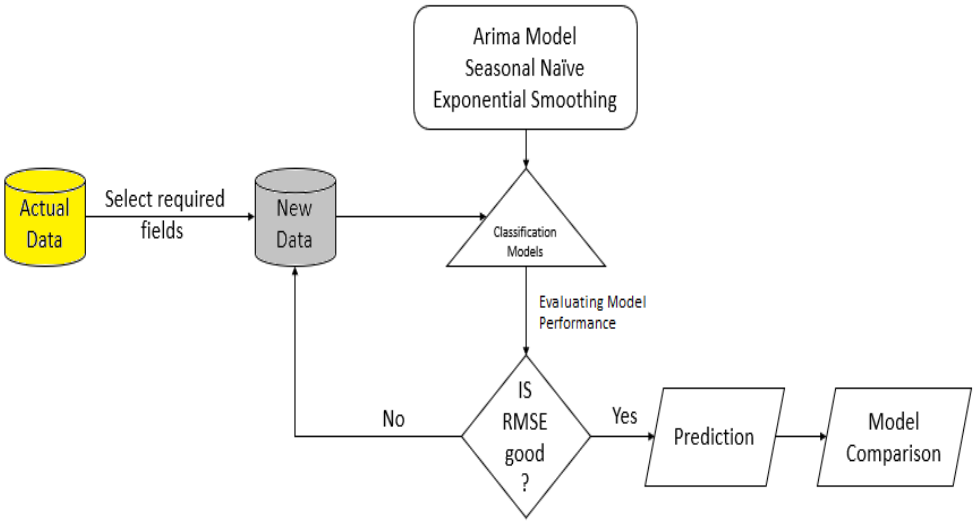

Fig.1 Flow diagram for sales prediction

Fig.1 shows the process flow sales prediction. The actual dataset (raw data) is explored and the attributes of datasets are filtered based on RMSE the level of requirements for the sales prediction. The attributes such as sales, Days of week, Store id etc., are filtered for the prediction process. Once data analysis is done, then the new datasets will be processed for the various classification models. The error factors such as value, MAPE value, p-value etc., are considered before making predictions, to increase the accuracy. If the error value is high, (i.e p-value greater than 0.05) then, the difference of the time series is again done to remove the effects of seasonality and trends. Finally, the prediction results are compared for all the models and based on the error factors, the best fit model having low error range and high accuracy is identified.

VI. SYSTEM OVERVIEW

6.1. PERFORMINGCLASSIFICATIONMODELS

6.1.1. ARIMA MODEL

Various arima model is used to process the time series data and to forecast the sales. The Arima model(1,1,0)(0,1,0) is identified as the best fit model based on AIC and BIC value.

The coefficients of the arima model is -0.4827 and 0.1862. From the observations, Arima Model (1,1,0)(0,1,0) has the values AIC = 416.55 and BIC = 418.64 which is better, compared to other models and the p, d and q values are calculated as 1,1 and 0 respectively. The lower AIC and BIC value, higher the accuracy of model.

The p value is calculated as 1 based on PACF, The q value is calculated as 0 based on ACF, which tells how many MA terms are required to remove any autocorrelation in the stationarized series.The d value is 1, i.e. the number of differencing required to make the time series stationary. The p-value is 0.03 which is less than 0.05, which indicates the strong evidence to null hypothesis, so that the null hypothesis is rejected.

6.2. SEASONAL NAÏVE MODEL

The error factors such as RMSE value 5108.624, MAPE value is 300.8545, MASE value 1 and p-value as 0.1098, which determines the accuracy is less for the predicted results. The residual value is 5234.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

The RMSE value is high compared to Arima model. By performing Ljung – box test, the p-value is observed as 0.1098 which is greater than the 0.05, which is not good enough for prediction.

6.3. EXPONENTIAL SMOOTHING

From observation, the alpha value is 1 e^{-04}. It uses single exponential smoothing model, which doesn't have any seasonality or trends. Thus, it uses only single parameter alpha.
The AIC and BIC value of ETS model is 663.9118 and 668.4909, which follows the ARIMA model (0,1,1). The p-value is observed as 0.3455 which indicates the accuracy is not good for prediction. The residual of ETS model is 2813.841.

6.4.PERFORMANCE EVALUATION

The error factors such as RMSE, MASE,MAPE, p-value, MPE etc., are used to determine the accuracy of each models.

	ME	RMSE	MAE	MPE	MAPE	MASE	P-value
ARIMA	8.89313	3485.91	2275.33	-4.7923	26.1046	0.772501	0.03545
SNAIVE	0.030061	2729.827	2126.27	-9.47093	25.93737	0.721893	0.1098
ETS	48.5714	5108.62	4149.42	62.9136	300.854	1	0.3455

Fig.2. Comparison of Error factors

Fig.2 shows the multiple error factors for all 3 models. Based on the error factors, the accuracy of the model is determined, and the best fit model is identified.

VII. RESULTS

We conduct our research in sales dataset which is converted in a timeseries format. Multiple machine learning models are applied, and the sales is forecasted for the upcoming 2 years. The forecasted sales are follows.

7.1. ARIMA MODEL

Residual = 4545
Mean Absolute Error (MAE) = 2275.332
Root Mean Square Error (RMSE) = 3485.919
Mean Percentage Error (MPE) = -4.792
P-value = 0.03545

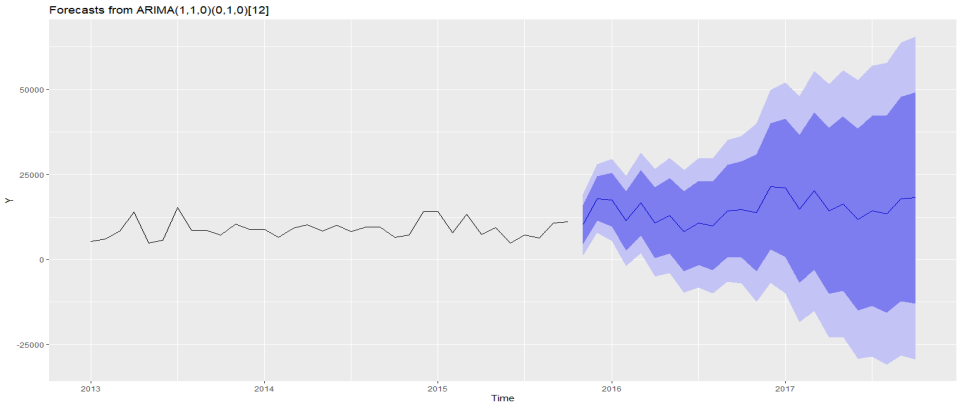

Fig.3 Arima Model Forecasted Data

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

Fig.3 shows the resultant Arima model's sales prediction for the upcoming years with the with the confidence level of 80% and 95%. The x-axis contains years and y-axis contains the sales.

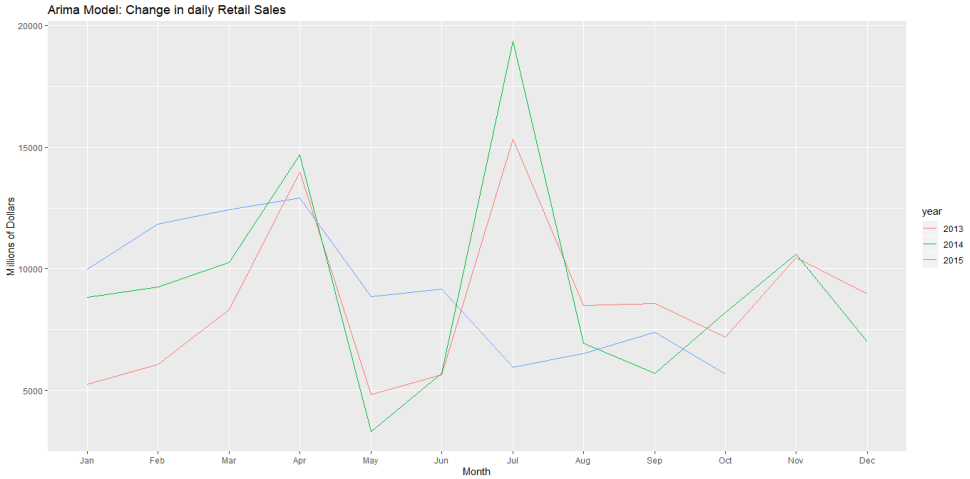

Fig.4 Data Variance in Sales Using Arima Model

Fig.4 shows the variation of sales across the years 2013,2014 and 2015. The graph shows the data variance, thus minimum data variance results in maximum accuracy of prediction. The x-axis contains months and y-axis contains the sales value per month.

7.2. SEASONAL NAÏVE MODEL

Residual = 5234
Mean Absolute Error (MAE) = 4149.429
Root Mean Square Error (RMSE) = 5108.624
Mean Percentage Error (MPE) = 62.9
P-value = 0.1098

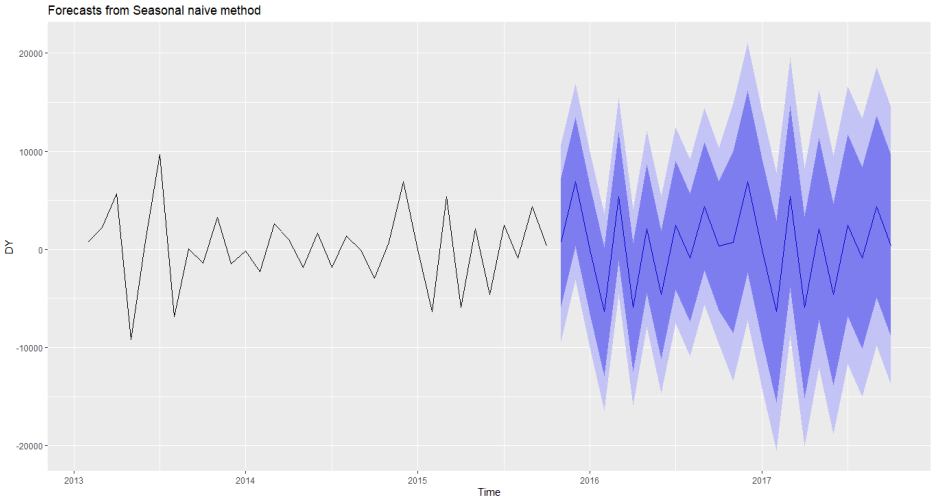

Fig.4 Seasonal Naïve Model Forecasted Data

Fig.4 shows the resultant Seasonal Naive model's sales prediction for the upcoming years with the with the confidence level of 80% and 95%. The x-axis contains years and y-axis contains the sales.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

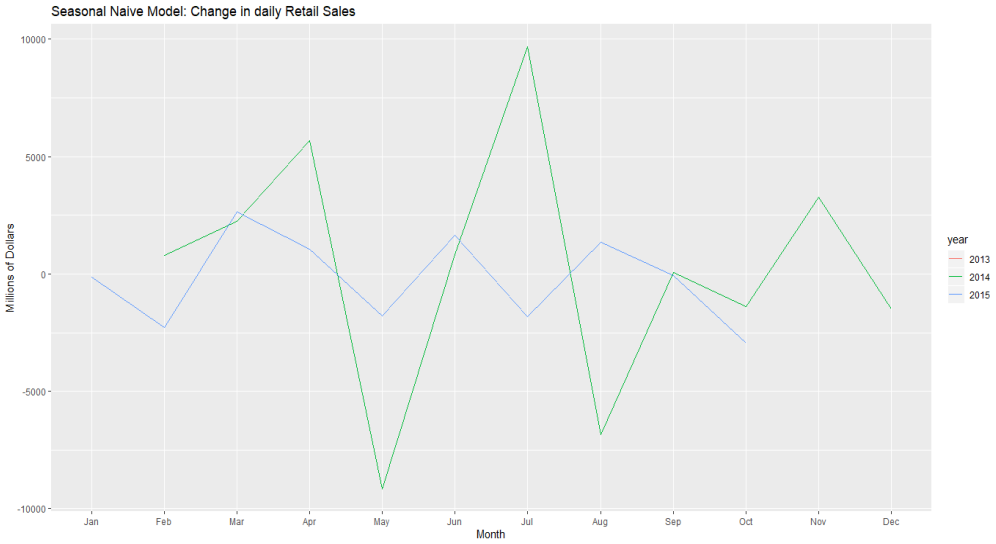

Fig.5 Data Variance in Sales Using Seasonal Naïve Model

Fig.5 shows the variation of sales across the years 2013,2014 and 2015. Here 12 rows containing missing values are removed. The x-axis contains the months and y-axis contains the sales.

7.3. EXPONENTIAL SMOOTHING MODEL

Residual = 2813
Mean Absolute Error (MAE) = 2126.27
Root Mean Square Error (RMSE) = 2729.827
Mean Percentage Error (MPE) = -9.47
P-value = 0.3455

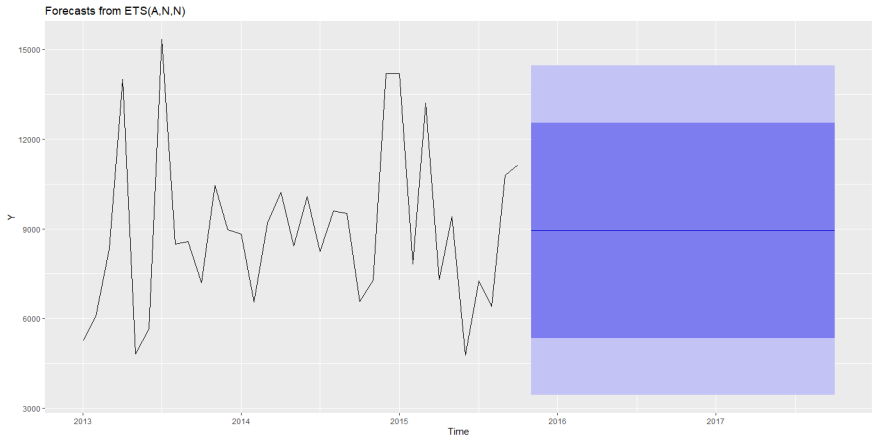

Figure.6 Exponential Smoothing Model Forecasted Data

Fig.6 shows the resultant ETS model's sales prediction for the upcoming years with the with the confidence level of 80% and 95%. The x-axis contains years and y-axis contains the sales.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

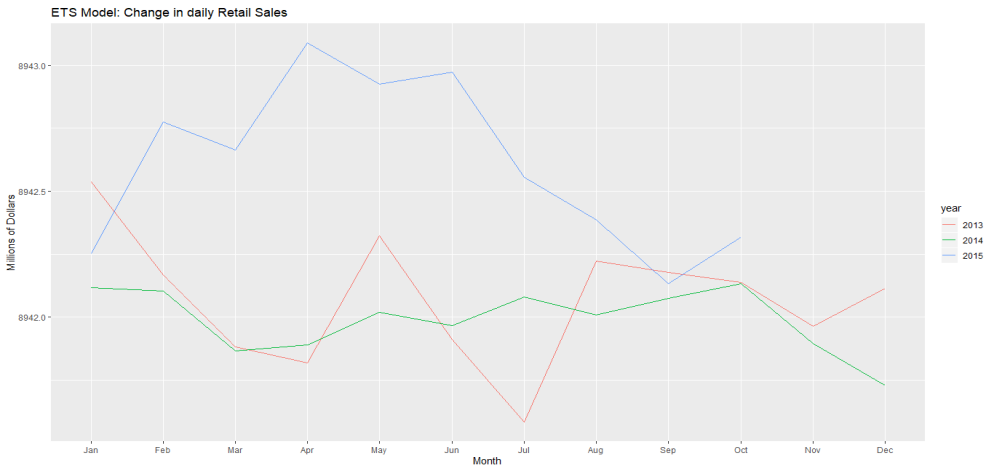

Figure.7 Data Variance in Sales Using Exponential Smoothing Model

Fig.7 shows the variation of sales across the years 2013, 2014, and 2015. Here the data variance is more compared to other 2 models, thus resulting in less accuracy in forecasting. The x-axis contains the months and y-axis contains the sales.

VIII. CONCLUSION AND FUTURE WORK

The algorithms aim is to develop the most accurate predictions for future sales with the current sales dataset. From the analysis and algorithms, the Arima model (1,1,0)(0,1,0) provides the best results in comparison with other model algorithms and the variance in data is less, so that the predicted results are more accurate. Arima model provides a p-value of 0.034 which is nearly equals to 0.05 which reduces the null hypothesis and it is statistically significant. It has Mean Absolute Percentage Error as 26% which is near to 25%, thus provides good accuracy and has Mean Absolute Scaled Error as 0.7 which increases the accuracy. It provides MASE value of 0.77 which is less than 1, thus increasing the accuracy. Thus, the Arima model (1,1,0) (0,1,0) provides overall accuracy of 74% of sales prediction by using daily sales data for the past 3 years.

Future work can be executed in continuing the study of Sales Prediction to improve the current algorithms. It can be concluded that, there are many factors that affects the sales forecasting. Factors such as past economic performance, current global conditions, rate of inflations, marketing efforts etc. need to be considered to forecast with more accuracy. To achieve this, correlations between different entities of Sales forecasting can be considered. Algorithms such as Random Forest, Linear regressions, KNN algorithm, K-means algorithm can be applied to forecast the sales. Approaching multiple algorithms will improve the accuracy and correctness of prediction sales.

References

[1] A.S. Harsoor and A. Patil, “Forecast of sales of Walmart store using Big data application”, International Journal of Research in Engineering and Technology, vol. 4, p. 6, June 2015.
[2] Manpreet Singh, BhawickGhutla, Reuben Lilo Jnr, Aeesan FS Mohammed, Mahmood A Rashid, “Walmart Sales Data Analysis- A Big Data Analytics Perspective”, 4th Asia-Pacific World Congress on Computer Science and Engineering, December 2017.
[3] SunithaCherian, Shaniba Ibrahim, SajuMohanan, Susan Treesa, “Intelligent Sales Prediction Using Machine Learning Techniques”, International Conference on Computing Electronics & Communication Engineering, August 2018.
[4] Bohdan M. Pavlyshenko, “Machine Learning Model For Sales Forecasting”, IEEE Second International Conference on Data Stream Mining & Processing, August 2018.
[5] Mentzer, J.T.; Moon, M.A. Sales Forecasting Management: A Demand Management Approach; Sage: Thousand Oaks, CA, USA, 2004.
[6] Efendigil, T.; Önüt, S.; Kahraman, C. A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. Expert Syst. Appl. 2009, 36, 6697–6707.
[7] Zhang, G.P. Neural Networks in Business Forecasting; IGI Global: Hershey, PA, USA, 2004.
Chatfield, C. Time-Series Forecasting; Chapman and Hall/CRC: Boca Raton, FL, USA, 2000.
[8] Brockwell, P.J.; Davis, R.A.; Calder, M.V. Introduction to Time Series and Forecasting; Springer: Cham, Switzerland, 2002; Volume 2.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 3928 - 3936
Received 15 April 2021; Accepted 05 May 2021.

[9] Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. Time Series Analysis: Forecasting and Control; John Wiley & Sons: Hoboken, NJ, USA, 2015.
[10] Doganis, P.; Alexandridis, A.; Patrinos, P.; Sarimveis, H. Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. J. Food Eng. 2006, 75, 196–204
[11] Hyndman, R.J.; Khandakar, Y. Automatic Time Series for Forecasting: The Forecast Package for R; Number 6/07; Monash University, Department of Econometrics and Business Statistics: Melbourne, Australia, 2007.