Prediction of Sales Value in online shopping using Linear Regression

Gopalakrishnan T [1]
Department of Computer Science and Engineering
School of Computing and IT,
Manipal University Jaipur, Rajasthan, India
gopalakrishnan.ct@gmail.com

Ritesh Choudhary[2]
Department of Information Technology
School of Computing and IT,
Manipal University Jaipur, Rajasthan, India
ritesh.choudhary456@gmail.com

Sarada Prasad[3]
Department of Information Technology
School of Computing and IT,
Manipal University Jaipur, Rajasthan, India
Sarada1987@gmail.com

***Abstract*—The aim of this paper is to analyze the sales of a big superstore, and predict their future sales for helping them to increase their profits and make their brand even better and competitive as per the market trends by generating customer satisfaction as well. The technique used for prediction of sales is the Linear Regression Algorithm, which is a famous algorithm in the field of Machine Learning. The sales data is from the year 2011-13 and prediction of data for the year 2014 is done. Then, real-time data of the year 2014 is also taken and the actual data of the year 2014 has been compared to the predicted data to calculate the accuracy of prediction. This is done so as to validate our results with the actual ones. This in turn would help them take necessary actions (which has been discussed later) for their increase their sales.**

Keywords— Data Visualization, Forecasting, Machine Learning, Regression, Linear Regression.

I. Introduction

We live in a world full of data. Data surrounds us everywhere. Right from handling monthly budgets, storing information on mobile phones, buying items from stores, all of it is stored in the form of data. In our everyday lives, we have to deal with a lot of data. This data could be as small as handling your monthly budgets to big ones like the data of a Multinational Company (often referred to as big data).

One of the agents who work with a lot of data is a superstore, like ZMart. These big shopping complexes have a lot of data to work upon. Handling inventory, maintaining purchase from manufacturers, handling inventory costs, handling supplies data, handling with their sales, profits and quantity data, and many more. This is a tremendous task to work upon such a big dataset. Our ultimate task is generating profits and customer satisfaction and to maintain brand name. A lot of work has to be done on the dataset for its analysis and prediction. This whole work is done so as to check the current position of sales and find out the future expected sales so that if any decline or anomalies is found could be worked upon by doing proper market research and evaluating the trends of the market so that customer base could be increased. Also, within the store, what techniques (like putting discounts or updating inventory) could be applied so that our target customers increase their purchase and become a satisfied and a happy customer. All this is very important for any business to survive in this cut-throat competition and undoubtedly data science is very much required to fulfill this purpose.

In this paper, we will describe the methodology to deal with such data along with predicting sales of the superstore for next years from the available tools like machine learning. A brief description about the processes involved in fulfilling our objective is discussed further. Also, the tools used for various processes would be even discussed.

As we know, first we define our data set and ask questions on what analysis or prediction are we looking from this data. After which we procure data from various sources and collect it. Next, the main difficult task is how to interpret information from such a large and unorganized data. Without proper processing the data, which is done by the process called as Data Architecture [2], we cannot do any further analysis on it. The process of analyzing data to infer valuable information from the dataset is called data analysis [3]. Data Analysis is a complete process which involves various steps in it right from organizing, cleansing and then obtaining results from it. Next, we move to the stage of Data Visualization [4]. In this step, we visualize the available processed datasets using various graphs and plots. This way we are able to infer a lot from the dataset. We could see the trend and could do get certain results which we were looking for. This process is called Data Analysis [5,10]. This way we could obtain valuable insights from the data. But, this is not enough. There is a possibility that from this analysis, we can certain more results from the available results, with the help of graphs and charts. This process is the crux of this paper and is called data analytics. This is achieved with the help of certain models and algorithms which are a part of machine learning. A detailed discussion on analytics would be done in this paper later. And finally, we predict accuracy of our predicted data set among various years to generate the accuracy rate

II. Research Methodology

The proposed approach was organized into three stages, first is data collection, which includes collecting data and transforming it into processed data. Then, it includes modelling the data for predictions using machine learning techniques (focusing on linear regression algorithm). And, finally validating and implementation of our results using precision and accuracy techniques. The complete cycle focusing on predictive analytics is shown in figure 1.

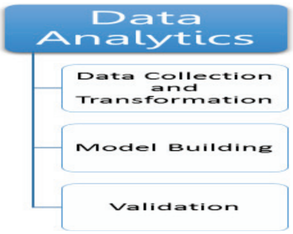


Fig. 1. Steps in Data Analytics Process

978-1-5386-6947-1/18/$31.00 ©2018 IEEE

As we can see in figure 1, the data is first accessed by the user from various sources. He explores the given dataset to find validate if it is ready for doing analytics over it. Mostly, the data isn't available in processed form so data transformation is used to cleanse the data and remove noises. Then, using appropriate modes, modelling is done on the dataset and results are calculated accordingly. And finally, the data is validated using precision and accuracy techniques and final result is implemented. A detailed discussion on each of the following steps will be done further.

A. Data Collection and transformation

Data is collected from various sources and organized in a single file. Next, the data cleansing process is applied. Data cleansing is the process of detecting and correcting inaccurate or obsolete records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Also, various noises within the data is also removed. Along with that, the data is categorized properly and all blank spaces or irrelevant information within data is also removed so that analytics could be performed easily on the data.

B. Model Building using Linear Regression

The processed data is used for predictive modelling so that appropriate results can be generated from it. This predictive modelling is done using a technique called Machine Learning [6]. It is defined as a "computer's ability to learn without being explicitly programmed". At its most basic, machine learning uses programmed algorithms that receive and analyse input data to predict output values within an acceptable range. As new data is fed to these algorithms, they learn and optimise their operations to improve performance, developing 'intelligence' over time. There are four types of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement. Out of which, we will look only on supervised and unsupervised learning in this paper.

Supervised learning- In supervised learning, the machine is taught with the help of examples. The user provides the ML algorithm with a dataset that includes desired inputs and outputs, and the algorithm finds a method to determine how to arrive at those results.

Unsupervised learning- Here, the ML algorithm examines data to identify patterns. There is no user to provide instructions. Instead, the machine determines the correlations and relationships by analyzing available data.

The most common and popular machine learning algorithms are-

- Naïve Bayes Classifier Algorithm (Supervised Learning - Classification)
- K Means Clustering Algorithm (Unsupervised Learning - Clustering)
- Support Vector Machine Algorithm (Supervised Learning - Classification)
- Linear Regression (Supervised Learning/Regression)
- Logistic Regression (Supervised learning – Classification)
- Decision Trees (Supervised Learning – Classification/Regression)
- Random Forests (Supervised Learning – Classification/Regression)
- Nearest Neighbours (Supervised Learning)

The predictive analytics in this paper is done using Linear Regression [7] algorithm, which will be discussed further now.

C. Linear Regression

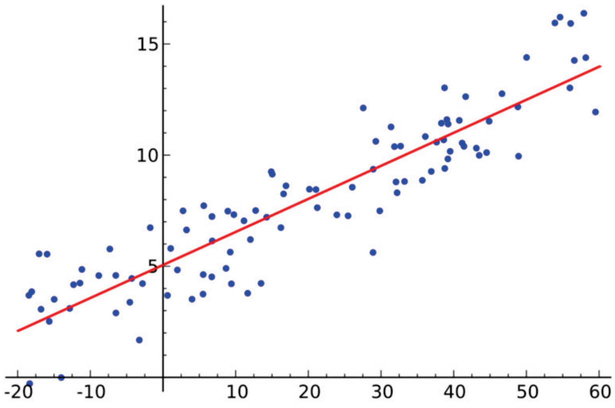


Fig. 2. Sample Linear Regression model for Students marks

Linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable [1]. The red line in the above graph is referred to as the best fit straight line. The line can be modelled based on the linear equation shown below.

y = a_0 + a_1 * x (1)

The motive of the linear regression algorithm is to find the best values for a_0 and a_1. Before moving on to the algorithm, let's have a look at two important concepts that must be known to better understand linear regression.

D. Cost Function

The cost function helps to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. Since the best values is desired for a_0 and a_1, the search problem is converted into a minimization problem where it is desired to minimize the error between the predicted value and the actual value.

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

E. Minimization and Cost Function

The above function to chosen to be minimized. The difference between the predicted values and ground truth measures the error difference. The error difference is squared and summed over all data points and that value is divided by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function. Now, using this MSE function the values of a_0 and a_1 gets changed such that the MSE value settles at the minima.

F. Gradient Descent

The next important concept needed to understand linear regression is gradient descent. Gradient descent is a method of updating a_0 and a_1 to reduce the cost function(MSE)[7]. The idea is to start with some values for a_0 and a_1 and then these values will be changed iteratively to reduce the cost. Gradient descent helps on how to change the values.

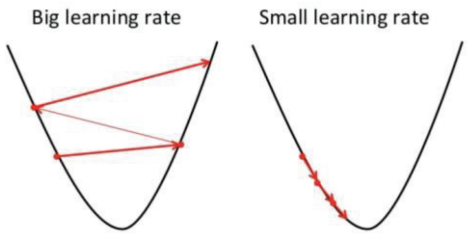


Fig. 3. Gradient Descent

To draw an analogy, imagine a pit in the shape of U and someone is standing at the topmost point in the pit and his objective is to reach the bottom of the pit. There is a catch, he can only take a discrete number of steps to reach the bottom. If he decides to take one step at a time he would eventually reach the bottom of the pit but this would take a longer time. If he chooses to take longer steps each time, he would reach sooner but, there is a chance that he could overshoot the bottom of the pit and not exactly at the bottom. In the gradient descent algorithm, the number of steps he takes is the learning rate. This decides on how fast the algorithm converges to the minima.

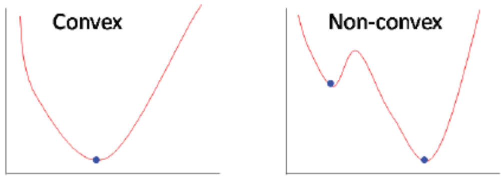


Fig. 4. Convex Vs Non-convex function

Sometimes the cost function can be a non-convex function where it tends to settle at local minima but for linear regression, it is always a convex function. Now, to update a_0 and a_1, gradients from the cost function is taken. To find these gradients, partial derivatives with respect to a_0 and a_1 is taken.

$$a_0 = a_0 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i)$$

$$a_1 = a_1 - \alpha \cdot \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i) \cdot x_i$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i - y_i)^2$$

$$\frac{\partial J}{\partial a_0} = \frac{2}{n} \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i - y_i) \implies \frac{\partial J}{\partial a_0} = \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i)$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n} \sum_{i=1}^{n} (a_0 + a_1 \cdot x_i - y_i) \cdot x_i \implies \frac{\partial J}{\partial a_1} = \frac{2}{n} \sum_{i=1}^{n} (pred_i - y_i) \cdot x_i$$

The partial derivate' s are the gradients and they are used to update the values of a_0 and a_1. Alpha is the learning rate which is a hyperparameter that must be specified. A smaller learning rate leads to getting closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that the minima could be overshoot.

III. Validation

To evaluate the efficiency of the proposed method, performance is measured in two factors namely precision [8-9]. Precision here specifies about the quantity of correct recommendations i.e. proportion of the relevant revivals to the total number of populations. Precision [8] is given by the following formula,

$$precision = \frac{T(p) \cap R(p)}{R(p)} \qquad (2)$$

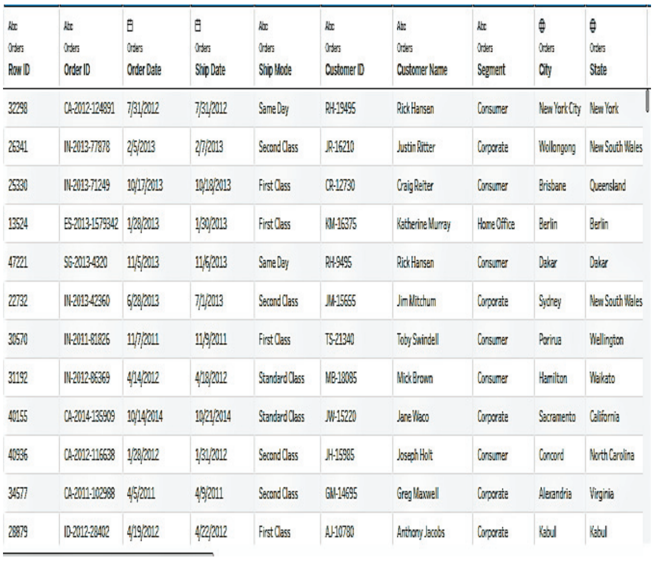
Where R(p) is recommendation set

T(p) is session

IV. Experimental Setup

In this paper, the dataset is collected from the various departments of ZMart superstore, including the finance, sales and marketing departments. Some inputs are also provided by the HR department as well. Apart from that, certain data is also retrieved from the web servers of ZMart by their permission to access. The data was collected and processed in Microsoft Excel. It took nearly 4 days to process this dataset for further use. The complete dataset post processing was stored in a Comma Separated Values (CSV) file. After removing all unwanted information, blank spaces and outliers, the dataset contains a total of 25 fields or columns. The dataset was now exported to another really powerful software named Tableau. Tableau is a software used for data visualization and further predictions and forecasting. The dataset after exporting into Tableau is shown in figure 5.

Next, the dataset is set for visualization. Our objective is to predict year's vs sales data. So, first, we need to plot quarters in a year vs sales data from the entire dataset in a graph. As it can be observed, quarters vs sales are a time-series data. A time-series data usually contains 0 or more dimensions and 1 or more measures. There are various graphs available in Tableau to plot such data. This includes heat maps, highlight tables, pie charts, bar graphs, tree maps, circle views, line and area chart. For simplicity and a better view, we are going to use a line chart for demonstration purpose. Fig 6 shows the plotting of a line chart of various quarters of years 2011-13.



Abc Orders Row ID	Abc Orders Order ID	Orders Order Date	Orders Ship Date	Abc Orders Ship Mode	Abc Orders Customer ID	Abc Orders Customer Name	Abc Orders Segment	Orders City	Orders State
32298	CA-2012-124891	7/31/2012	7/31/2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York
26341	IN-2013-77878	2/5/2013	2/7/2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South Wales
25330	IN-2013-71249	10/17/2013	10/18/2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland
13524	ES-2013-1579342	1/28/2013	1/30/2013	First Class	KM-16375	Katherine Murray	Home Office	Berlin	Berlin
47221	SG-2013-4320	11/5/2013	11/6/2013	Same Day	RH-9495	Rick Hansen	Consumer	Dakar	Dakar
22732	IN-2013-42360	6/28/2013	7/1/2013	Second Class	JM-15655	Jim Mitchum	Corporate	Sydney	New South Wales
30570	IN-2011-81826	11/7/2011	11/9/2011	First Class	TS-21340	Toby Swindell	Consumer	Porirua	Wellington
31192	IN-2012-86369	4/14/2012	4/18/2012	Standard Class	MB-18085	Mick Brown	Consumer	Hamilton	Waikato
40155	CA-2014-135909	10/14/2014	10/21/2014	Standard Class	JW-15220	Jane Waco	Corporate	Sacramento	California
40936	CA-2012-116638	1/28/2012	1/31/2012	Second Class	JH-15985	Joseph Holt	Consumer	Concord	North Carolina
34577	CA-2011-102988	4/5/2011	4/9/2011	Second Class	GM-14695	Greg Maxwell	Corporate	Alexandria	Virginia
28879	ID-2012-28402	4/19/2012	4/22/2012	First Class	AJ-10780	Anthony Jacobs	Corporate	Kabul	Kabul

Fig. 5. Data represented in Tableau

In the above representation, a line chart is plotted containing 12 Quarters (Q1 2011-Q4 2014) along the columns or the x-axis and sum of sales of individual entities in each quarter in the rows or the y-axis. As we can see, the sales value is nearly more than 1000K, precisely it is 1,072,850. Now, we will predict the sales values for the next 4 quarters of 2014. We will use the Linear Regression technique to do this prediction or forecasting, and show the forecasted results in another line chart as shown in figure 7.

As it can be observed in figure 7, the highlighted blue part in the right end of the chart is showing the forecasted sales value for the 4 quarters of 2014. The line in between which is continuation to line from 2011-13 depicts the average value of sales of the 2014, while the highlighted part above and below the blue line shows the deviation from the line or simply the range in which the actual sales value might lie in reality during the year 2014. As it can be inferred, the range between which the sales value would keep on fluctuating amongst the 4 quarters lies between 600K to 1300K (Q1-014 to Q4-2014). The average sales value throughout the year, depicted by the blue line is 1,241,699.

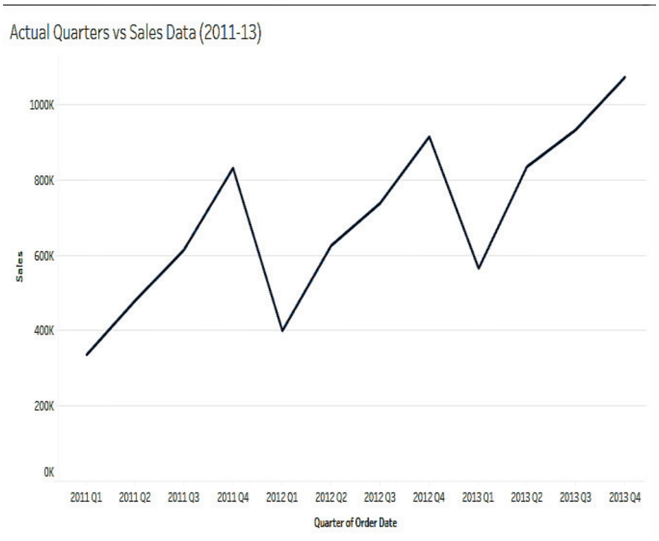


Fig. 6. Representation of quarters vs sales data using a line chart

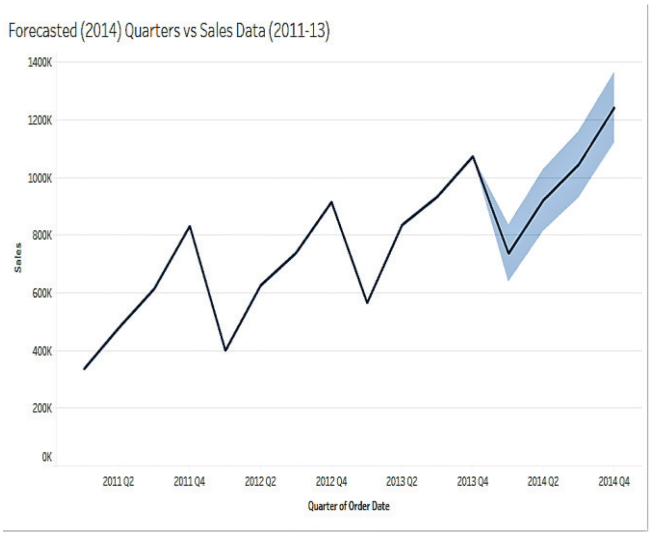


Fig. 7. Forecasted Quarters vs Sales data (2011-14)

V. Results

As it was observed, our predicted sales value for the 4 quarters of 2014 was 1,241,699. Our next task is to measure the accuracy rate of this result. We know that the data we are working upon is a time-series data and is a real-time dataset. Now, we've extracted the actual sales value of the year 2014 as well from the ZMart executives. From this data, we will try to compare the accuracy of our prediction. We will check the accuracy of our prediction of sales value of 2014 to the actual sales value of the year 2014. Fig-8 shows the actual data of the year 2014 and fog-9 shows the comparison of actual vs predicted sales.

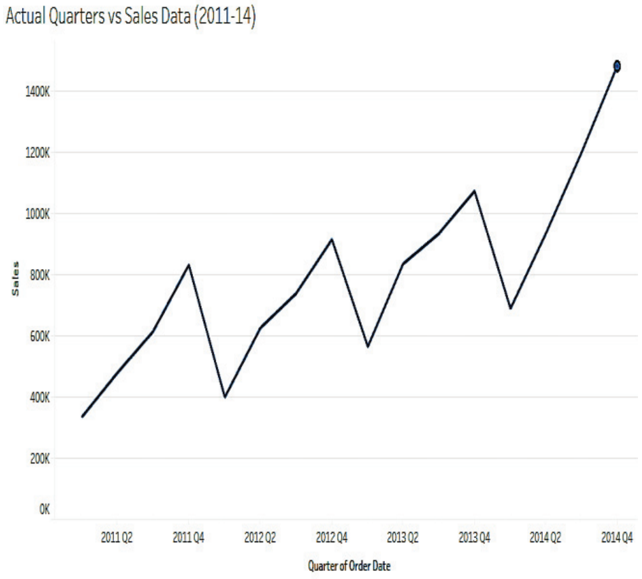


Fig. 8: Actual Quarters vs Sales Data for the years 2011-14

As we can see in Fig-8, the actual sales value after Q4 2014 is somewhat near to 1500K. To be precise, it is 1,481,189, which shows quite an upper inclination after the year 2013. Now, in Fig-9, we've created a dashboard [9], which in we have shown two sheets together, Actual Quarters Vs Sales Data (2011-14) and Forecasted Quarters Vs Sales Data (2011-14), for comparing the two. As we can observe, the shaded blue region which depicts the range of prediction within which the actual sales value after Q4 2014 is very close to the actual sales after Q4 2014, making our prediction level very good. But, our main focus lies with the blue line which shows the average sales value of 2014, this is because our objective is to measure the accuracy rate of our prediction with the actual data. To achieve this purpose, we have to perform some simple calculations.

We know,
Actual Sales Value after Q4 2014 = 1,481,189
Predicted Sales Value after Q4 2014 = 1,241,699

Let us measure the percentage error between the actual sales and predicted sales after Q4 2014:

$$\text{Percentage Error} = \frac{1{,}481{,}189 - 1{,}241{,}699 \quad * 100}{1{,}481{,}189}$$

$$= 16.168\ \%$$

To measure the accuracy rate, we subtract the percentage error from 100. So, we get:

$$\text{Percentage Accuracy} = (100 - 16.168)\ \%$$

$$= 83.832\ \% \sim 84\%$$

So, after calculations, we observe our accuracy rate to be 84% approximately

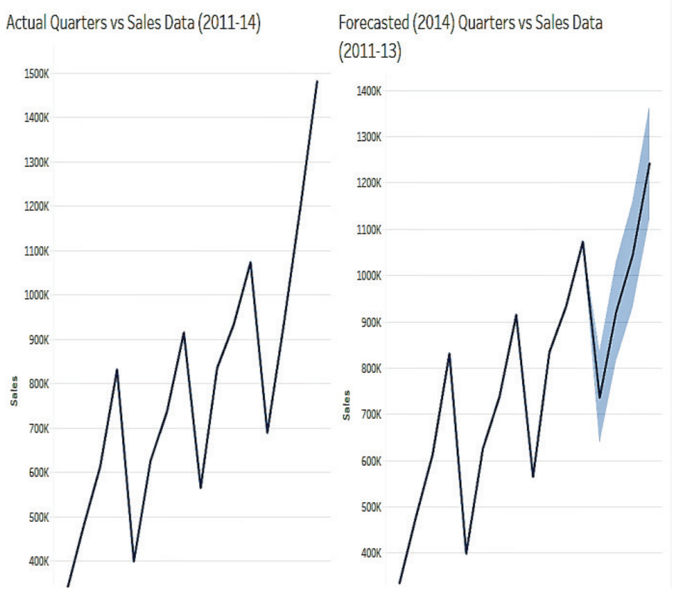


Fig. 9: Dashboard containing comparison of actual vs predicted quarters vs sales data for the years 2011-14

VI. CONCLUSION

Thus the main objective to predict the sales for the year 2014 from the given data of the years 2011-13 is performed. We also compared the accuracy rate from the actual data of the year 2014 to the predicted data of the year 2014, and we got the accuracy rate to be 84%. The purpose of measuring accuracy was to validate our prediction with the actual result. This is an important step to generate trust of target audience so that they can believe to our predictions are correct and take necessary actions accordingly. Our target audience were the ZMart executives. We forecasted their sales from their given data of the years 2011-13 for the year 2014. But, for them to believe whether our predictions are right or wrong, we took actual data from them for the year 2014 and then compared that with our forecasted sales value and found our predictions 84% accurate, which indeed is very close. With this, the sales can also be predicted for the year 2015, and results can be shown. With our predictions, they can refine their methodologies and strategies to increase their sales of products.

REFERENCES

[1] "Applied Linear Statistical Models" Fifth Edition by Kutner, Nachtsheim, Neter and L, Mc Graw Hill India, 2013, Paperback, 9781259064746

[2]]. Demchenko, Yuri & de Laat, Cees & Membrey, Peter. (2014). Defining architecture components of the Big Data Ecosystem. 104-112. 10.1109/CTS.2014.6867550.

[3] https://www.bigskyassociates.com/blog/bid/372186/The-Data-Analysis-Process-5-Steps-To-Better-Decision-Making

[4] Stevencua, Stevencua & Setiawan, Johan. (2018). Data Visualization of Poverty Level at Provinces in Indonesia from The year 2013-2015. International Journal of New Media Technology. 5. 8-12. 10.31937/ijnmt.v5i1.813.

[5] Luigi Da, Immanuel & Setiawan, Johan. (2017). Data Visualization Indicator Disease (Malaria, Dengue Fever, and Measles) in The Year 2012-2015.

[6] https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html

[7] https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

[8] Alper Sarikaya, Michael Correll, Lyn Bartram, Melanie Tory, and Danyel Fisher. (2018). What Do We Talk About When We Talk About Dashboards?. IEEE InfoVis 2018 (Berlin, Oct 21-26, 2018)

[9] https://www.decisionanalyst.com/media/downloads/ConsumerDecisionMaking.pdf

[10] Sengottuvelan P & Gopalakrishnan T (2016). A hybrid PSO with Naïve Bayes classifier for disengagement detection in online learning, Program, ISSN: 0033-0337, vol. 50, no. 2, pp. 215-224