**Business Intelligence Lab**

**Experiment 02**

**Aim:**

To identify Business aspects for a identified domain and perform analysis for the same

**Theory:**

**Case study on Walmart with Big Data Real Time Analysis**

Walmart has a broad big data ecosystem. The big data ecosystem at Walmart processes multiple Terabytes of new data and petabytes of historical data every day. The analysis covers millions of products and 100’s of millions of customers from different sources.

The analytics systems at Walmart analyse close to 100 million keywords on daily basis to optimise the bidding of each keyword.The main objective of leveraging big data at Walmart is to optimise the shopping experience for customers when they are in a Walmart store, or browsing the Walmart website or browsing through mobile devices when they are in motion. Big data solutions at Walmart are developed with the intent of redesigning global websites.

Walmart uses data mining to discover patterns in point of sales data. Data mining helps Walmart find patterns that can be used to provide product recommendations to users based on which products were bought together or which products were bought before the purchase of a particular product. Effective data mining at Walmart has increased its conversion rate of customers.

Walmart tracks and targets every consumer individually.

Walmart has exhaustive customer data of close to 145 million Americans of which 60% of the data is of U.S adults. Walmart gathers information on what customer’s buy, where they live and what are the products they like through in-store Wi-Fi.The big data team at Walmart Labs analyses every clickable action on Walmart.com-what consumers buy in-store and online, what is trending on Twitter, local events such as San Francisco giants winning the World Series, how local weather deviations affect the buying patterns, etc.

All the events are captured and analysed intelligently by big data algorithms to discern meaningful big data insights for the millions of customers to enjoy a personalised shopping experience

**Dataset Description**

Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 Walmart stores. The business is facing a challenge due to unforeseen demands and runs out of stock some times, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modelling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data. Historical sales data for 45 Walmart stores located in different regions are available.

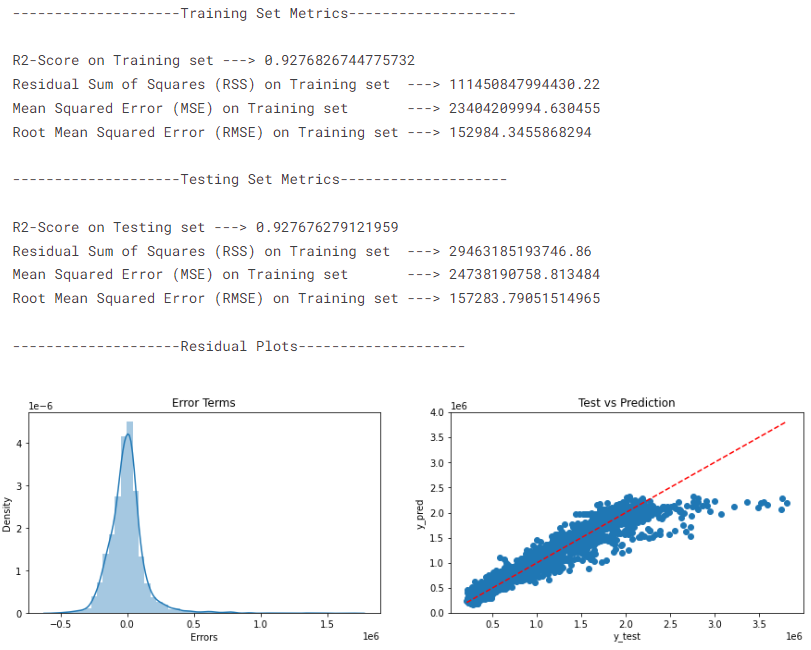
Dataset Info: This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in the file Walmart\_Store\_sales. Within this file you will find the following fields:

1. Store - the store number
2. Date - the week of sales
3. Weekly\_Sales - sales for the given store
4. Holiday\_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
5. Temperature - Temperature on the day of sale
6. Fuel\_Price - Cost of fuel in the region
7. CPI – Prevailing consumer price index
8. Unemployment - Prevailing unemployment rate
9. Holiday Events - Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13\ Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13\ Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13\ Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

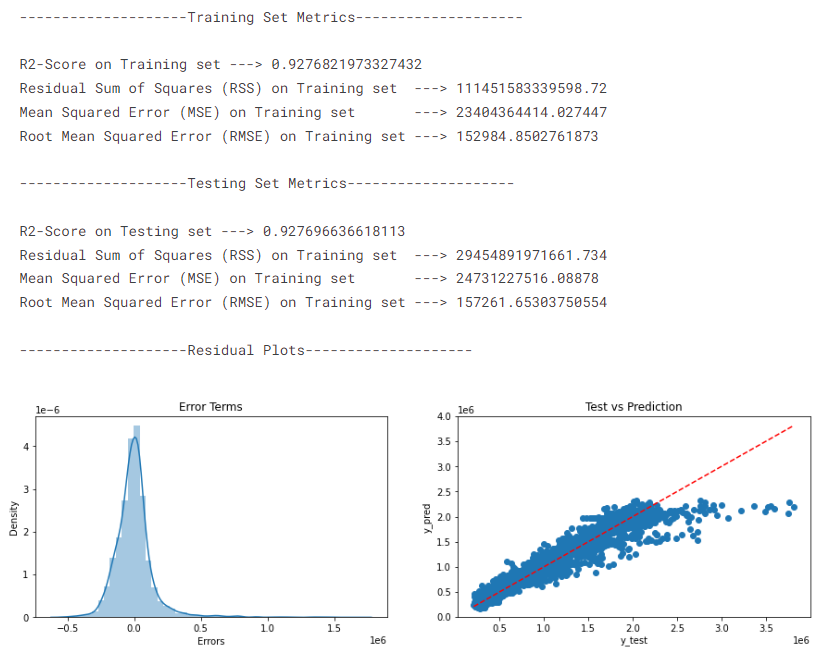
**Predictive Analysis**

Multiple regression models are used and most accurate model is used for prediction

1. Multiple Linear Regression
   1. Multiple regression is a statistical technique that can be used to analyse the relationship between a single dependent variable and several independent variables.
   2. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.
   3. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.
   4. Prediction of sales using this model is



1. Ridge Regression Model
   1. Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
   2. Ridge regression is a regularisation technique, which is used to reduce the complexity of the model. It is also called L2 regularisation.
   3. In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.
   4. Prediction of sales using this model is



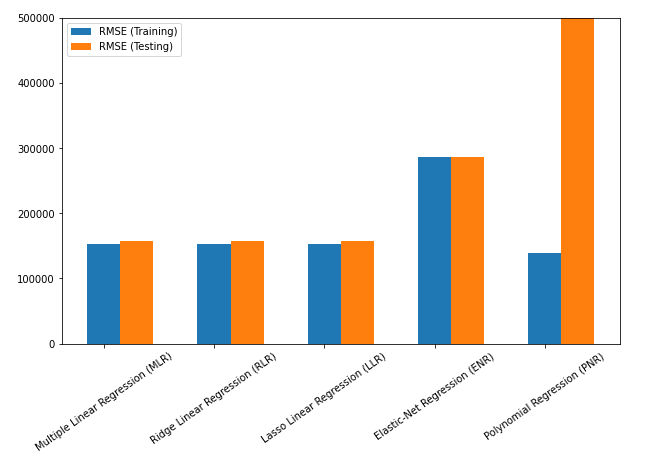
1. Elastic-Net Regression
   1. Elastic net is a penalised linear regression model that includes both the L1 and L2 penalties during training.
   2. Using the terminology from “The Elements of Statistical Learning,” a hyperparameter “alpha” is provided to assign how much weight is given to each of the L1 and L2 penalties. Alpha is a value between 0 and 1 and is used to weight the contribution of the L1 penalty and one minus the alpha value is used to weight the L2 penalty.
   3. elastic\_net\_penalty = (alpha \* l1\_penalty) + ((1 – alpha) \* l2\_penalty)
   4. For example, an alpha of 0.5 would provide a 50 percent contribution of each penalty to the loss function. An alpha value of 0 gives all weight to the L2 penalty and a value of 1 gives all weight to the L1 penalty.



1. Polynomial Regression Model
   1. Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial
   2. It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
   3. It is a linear model with some modification in order to increase the accuracy.
   4. The dataset used in Polynomial regression for training is of non-linear nature.
   5. It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.
   6. Hence, "In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modeled using a linear model."
   7. Taking 2nd order polynomial regression



After comparing all of the predictions, it is clear that the polynomial regression models have the highest accuracy in prediction.

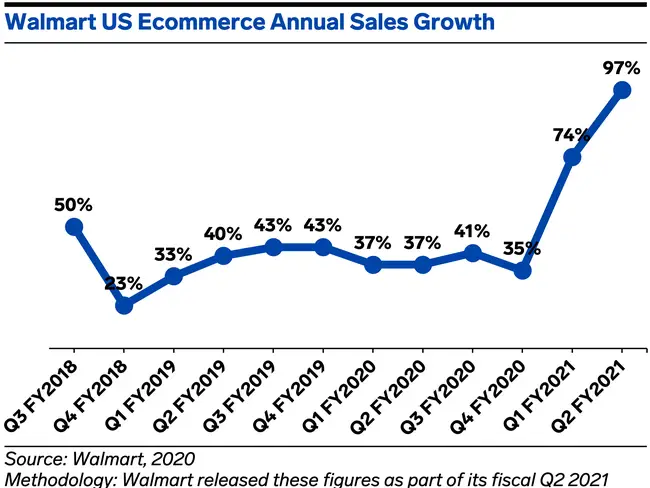


**Effect of these data analytics on Walmart:**

The Data Café allows huge volumes of internal and external data, including 40 petabytes of recent transactional data, to be rapidly modelled, manipulated and visualised. Senior statistical analyst Naveen Peddamail said, “If you can’t get insights until you’ve analysed your sales for a week or a month, then you’ve lost sales within that time.”

Quick access to insights is therefore vital. For example, Peddamail told about a grocery team who could not understand why sales had suddenly declined in a particular product category. By drilling into the data, they were quickly able to see that pricing miscalculations had led to the products being listed at a higher price than they should have been.

The system also provides automated alerts, so, when particular metrics fall below a set threshold in any department, the relevant team is alerted so that they can find a fast solution. In one example of this, during Halloween, sales analysts were able to see in real time that, although a particular novelty cookie was very popular in most stores, it wasn’t selling at all in two stores. The alert prompted a quick investigation, which showed that, due to a simple stocking oversight, the cookies hadn’t been put on the shelves. The store was then able to rectify the situation immediately.

****

**Conclusion**

In conclusion, the Walmart case study helped in understanding the need of different regression and how it helped Wal-Mart to become the number one retailer in the USA. There are other companies who are constantly rising as well and would give Walmart a tough competition in the future if Walmart does not stay at the top of their game. In order to do so, they will need to understand their business trends, the customer needs and manage the resources wisely.

In this era when the technologies are reaching out to new levels, Big Data is taking over the traditional method of managing and analysing data. These technologies are constantly used to understand complex datasets in a matter of time with beautiful visual representations. Through observing the history of the company’s datasets, clearer ideas on the sales for the previous years were realised which will be very helpful to the company on its own.

Additionally, seasonality trend and randomness and future forecasts will help to analyse sale drops which the companies can avoid by using a more focused and efficient tactics to minimise the sale drop and maximise the profit and remain in competition