Experiment No 02

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

Dataset: Big Mart Sales

Research Paper: Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms

Theory:
Case Study on Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms.

The constant competition between vibrant malls and enormous stores is becoming increasingly aggressive, and this violence is solely attributable to the rapid growth of both online and international promenades. The harshness and acrimony of the competition between several shopping malls and enormous supermarkets have increased as a result of the rise of international malls and internet shopping. To effectively draw a large number of customers and determine the number of sales for each product, as well as for the business' logistics, distribution, and stock management needs, each request seeks to offer substantiated and limited-time deals to attract numerous guests counting on a period of time. This allows each item's volume of deals to be estimated for the association's stock control, transportation, and logistical services.In order to outperform low-cost prediction techniques, the present machine learning is highly complex and gives prospects for forecasting or forecasting demand for any sort of company. Projections that are periodically updated are essential for developing and improving marketing plans that are relevant to particular markets.

Constantly improving vaccination is beneficial for creating and refining business marketing tactics, which is very beneficial. Nevertheless, not all machine learning methods are created equal, nor are they all equally accurate. A machine-learning method may therefore be extremely effective when used to solve one problem but worthless when used to solve another.Big Mart must therefore combine various machine-learning techniques in order to create a useful forecasting model. analytics-based revenue projections. Finding the most effective predictive analytics For Big Mart, we developed a functional prototype of a system for forecasting sales using machine learning. Before releasing this prototype, the algorithm needs to be tested on Big Mart. real information from Mart. As a result, we put our prototype to the test using sales data from Big Mart, and we built a machine-learning classifier model using two versions.

Breakdown of the Problem Statement: This is a supervised machine learning problem with a target label as (Item_Outlet_Sales). Also since we are expected to predict the sale price for a given product, it becomes a regression task.

Dataset Description:

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and predict the sales of each product at a particular outlet. Using this model, BigMart will try to understand the properties of products and outlets which play a key role in increasing sales. Please note that the data may have missing values as some stores might not report all the data due to technical glitches. Hence, it will be required to treat them accordingly. Within this file you will find the following fields.

1. Item_Identifier:- Unique product ID
2. Item_Weight:- Weight of product
3. Item_Fat_Content:- Whether the product is low fat or not
4. Item_Visibility:- The % of total display area of all products in a store allocated to the particular product
5. Item_Type:- The category to which the product belongs
6. Item_MRP:- Maximum Retail Price (list price) of the product
7. Outlet_Identifier:- Unique store ID
8. Outlet_Establishment_Year:- The year in which store was established
9. Outlet_Size:- The size of the store in terms of ground area covered 10.Outlet_Location_Type:- 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.Item_Outlet_Sales:- Sales of the product in the particular store. This is the outcome variable to be predicted.

Target/Dependent attributes

1. Item_Outlet_Sales

Input/Independent attributes

1. Item_Identifier
2. Item_Weight
3. Item_Fat_Content
4. Item_Visibility
5. Item_Type
6. Item_MRP
7. Outlet_Establishment_Year
8. Outlet_Size
9. Outlet_Location_Type
10. Outlet_Type

Types of attributes (Nominal, Ordinal, Continuou, Discrete)

● **Nominal**:-
1. Item_Identifier
2. Item_Fat_Content
3. Item_Type
4. Outlet_Identifier

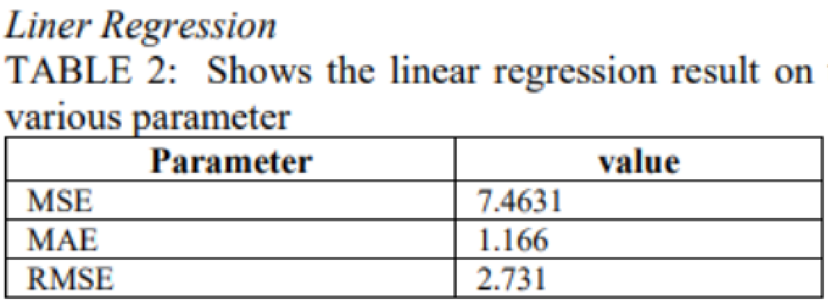
● **Ordinal**:-
1. Outlet_Size
2. Outlet_Location_Type
3. Outlet_Type

● **Continuous**
1. Item_Weight
2. Item_Visibility
3. Item_Outlet_Sales
4. Item_MRP

● **Discrete**
1. Outlet_Establishment_Year

Predictive Analysis:
After completing Data Preprocessing and Feature Transformation, the dataset is now ready to build a predictive model. The algorithm is fed into the training set in order to learn how to forecast values. After Model Building a target variable to forecast, testing data is supplied as input. The predictive models are built using

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 and the normal probability plot .A transformation might be necessary if the assumptions made do not appear to be met. Linear regression formulas look like this:
Y=o1x1+ o2x2+……… onxn



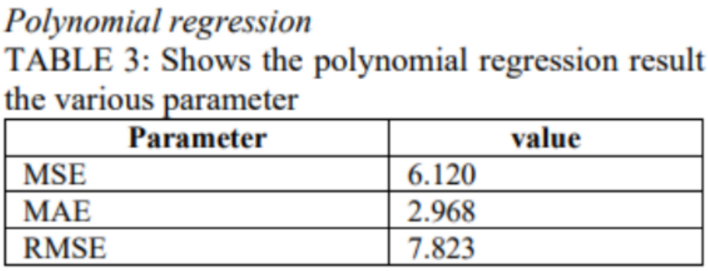
Liner Regression

TABLE 2: Shows the linear regression result on the various parameter

Parameter	value
MSE	7.4631
MAE	1.166
RMSE	2.731

B. Polynomial Regression Algorithm: Polynomial Regression is a relapse calculation that models the relationship here among dependent(y) and the autonomous variable(x) in light of the fact that as the most extreme limit polynomial. The condition for polynomial relapse is given beneath:

y= b0+b1x1+ b2x12+ b2x13+...... bnx1n



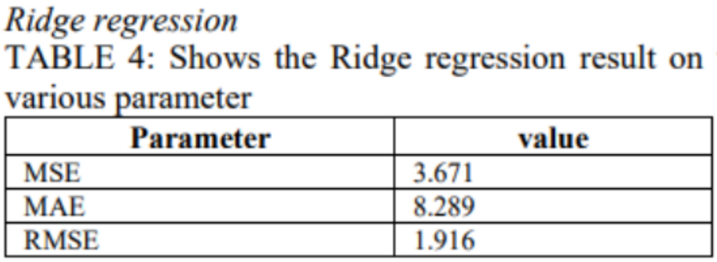
Polynomial regression

TABLE 3: Shows the polynomial regression result on the various parameter

Parameter	value
MSE	6.120
MAE	2.968
RMSE	7.823

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. Min(||Y – X(theta)||^2 + λ||theta||^2) The usual regression equation forms the base which is written as:

Y = XB + e

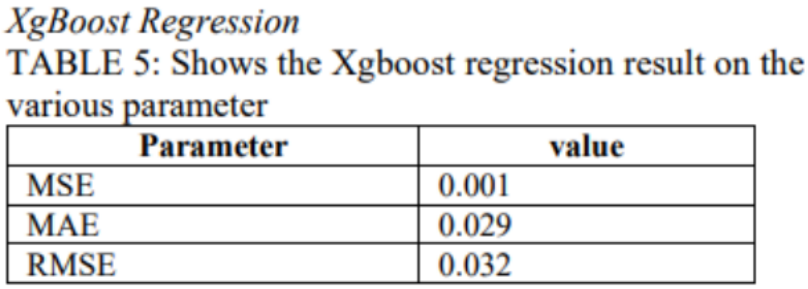


Ridge regression

TABLE 4: Shows the Ridge regression result on the various parameter

Parameter	value
MSE	3.671
MAE	8.289
RMSE	1.916

D. XGBoost: Extreme Gradient Boosting is the 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.



XgBoost Regression

TABLE 5: Shows the Xgboost regression result on the various parameter

Parameter	value
MSE	0.001
MAE	0.029
RMSE	0.032

Inferences and effects on business/value

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

Conclusion: The most effective algorithm uses a retrogression technique to forecast deals concentrating on actual deal data after analyzing the effectiveness of colorful algorithms on profit data. Prognostications may be more accurate when employing direct retrogression because of this method. Retrogressions along ridges and lines are also present. The Ridge, MAE, RMSE, and MSE retrogression approaches are the most successful, we can infer. There are two retrogression patterns related to vaticination perfection: direct and linear. It will be simpler to handle thanks to staffing, financial requirements, and transaction foretelling. planning a business. Future research using the ARIMA simulation may potentially make advantage of the time series graph, which displays data across time.