Experiment 2

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.

Everyday competitiveness between various shopping centers and huge marts is becoming more intense, violent just because of the quick development of global malls and online shopping. Each market seeks to offer personalized and limited-time deals to attract many clients relying on a period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services. The current machine learning algorithm is very advanced and provides methods for predicting or forecasting sales of any kind of organization, extremely beneficial to overcome low – priced used for prediction. Always better prediction is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful.

Bigmart is a big supermarket chain, with stores all around the country. The management of the shop had set out a challenge to all Data Scientists to help them create a model that can predict the sales per product for each store. The shop has collected sales data of products across 10 stores in different cities over a given period of time.

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:

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

1. **Linear Regression:**
    a. Linear regression analysis is used to predict the value of a variable based on the value of another variable. It is mostly used for finding out the relationship between variables and forecasting.
    b. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
    c. Linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.
    d. If the goal is error reduction in prediction or forecasting, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables.
    e. If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables

f. Prediction of sales using this model is

Liner Regression

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

Parameter	value
MSE	7.4631
MAE	1.166
RMSE	2.731

2. <u>Polynomial Regression:</u>

a. Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial
b. 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.
c. It is a linear model with some modification in order to increase the accuracy.
d. The dataset used in Polynomial regression for training is of non-linear nature.
e. It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.
f. 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"
g. Prediction of sales using this model is

Polynomial regression

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

Parameter	value
MSE	6.120
MAE	2.968
RMSE	7.823

3. **Ridge Regression:**
    a. 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.
    b. Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called L2 regularization.
    c. 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.
    d. Prediction of sales using this model is

Ridge regression

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

Parameter	value
MSE	3.671
MAE	8.289
RMSE	1.916

4. **XgBoost Regression:**
    a. Extreme Gradient Boosting (XgBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm.
    b. Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. XGBoost can be used directly for regression predictive modeling.
    c. The objective function of XgBoost contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values.
    d. The most common loss functions in XGBoost for regression problems is reg:linear, and that for binary classification is reg:logistics.
    e. Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XgBoost is one of the ensemble learning methods.
    f. XgBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancel out and better one sums up to form final good predictions.

g. Prediction of sales using this model is

XgBoost Regression

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

Parameter	value
MSE	0.001
MAE	0.029
RMSE	0.032

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

TABLE 7: Comparison of MAE, MSE, RMSE with the Model

Model	MSE	MAE	RMSE
Linear Regression	7.4631	1.166	2.731
Polynomial Regression	2.0364	7.002	1.427
Ridge Regression	3.6712	8.289	1.916
Xgboost Regression	0.001	0.029	0.0321

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:

In conclusion, the BigMart case study helped in understanding the need of different regression and how it helped BigMart to become successful in their field. There are other companies who are constantly rising as well and would give BigMart a tough competition in the future if BigMart 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.