Prepared by : Kaushik Prasad Dey, DBI\_0015

**ABSTRACT**:

**Machine Learning** is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization’s target.

Nowadays shopping malls and Big Marts keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the data store from the data warehouse.

The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart. In this paper, we propose a predictive model using **Linear, lasso, Ridge, Decision-tree, Random-forest, XG boost** Regressor technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models.

**Keywords**: Machine Learning, Sales Prediction, Big Mart, Linear Regression, Lasso regression, Ridge Regression, Decision-tree, Random Forest.

**Tools Used**

Icon

Description automatically generated A picture containing chart

Description automatically generated A picture containing text, clipart

Description automatically generated Chart, radar chart

Description automatically generated

**INTRODUCTION:**

In today’s modern world, huge shopping centers such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in prediction of future demands and inventory management.

Big Mart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a model that can predict the sales, per product, for each store to give accurate results.

The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse.

The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results that shed a new light on our knowledge with respect to the task’s data.

**PROBLEM STATEMENT:**

The Data scientist has collected sales data from **the year 2018**, for **1531** products across **8** stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.

Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

**Hypothesis Generation :**

Make it a practice to do this before solving any ML problem. Ideally, do it before seeing the data or else you might end up with biased hypotheses.

What could affect the target variable (sales)?

1. Time of week - Weekends usually are more busy.
2. Higher sales according to time of the day (morning and late evening).
3. Higher sales during end of the year.
4. Store size and location.
5. Items with more shelf space sell more.

**Methodology:**

The steps followed in this work, right from the dataset preparation to obtaining results are represented in Fig.1.

Implementing Machine learning Model

Testing & Choosing Best

Training of dataset

Pre-Processing

Raw Data

**PHASES IN MODEL :**

1. **Dataset :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Type | Subtype | Description | Segment | Expectation |
| Item Identifier | numeric | Discrete | Unique product ID | product | Low impact |
| Item\_Weight | numeric | Continuous | Weight of product | product | Medium impact |
| Item\_Fat\_Content | Categorical | Ordinal | Checks the Concentration of fat in the product | product | Medium impact |
| Item\_Visibility | numeric | Continuous | The % of total display area of all similar products in a store | product | High Impact |
| Item\_Type | Categorical | Nominal | Product Category | Product | High Impact |
| Item\_MRP | numeric | Discrete | Maximum Retail Price for a Product | Product | Medium impact |
| Outlet\_Identifier | numeric | Discrete | Unique Store ID | Store | Low impact |
| Outlet\_Establishment\_Year | numeric | Discrete | The year in which store was established | Store | Low impact |
| Outlet\_Size | Categorical | Ordinal | The size of the store (Area Size Category) | Store | High Impact |
| Outlet\_Location\_Type | Categorical | Ordinal | In Terms of city Tiers (Size) | Store | High Impact |
| Outlet\_Type | Categorical | Ordinal | Grocery store or a type of supermarket | Store | High Impact |
| Item\_Outlet\_Sales | numeric | Discrete | Sales of the product In the Specific outlet | Product | Target |

1. **Dataset and its Preprocessing :**

Big Mart’s data scientists collected sales data of their 8 stores situated at different locations with each store having 1531 different products as per 2018 data collection. Using all the observations it is inferred what role certain properties of an item play and how they affect their sales. The dataset looks like below on using head() function on the dataset variable.



Graphical user interface

Description automatically generated with medium confidence

The data set consists of various data types from integer to float to object as shown in below.

Text

Description automatically generated with medium confidence

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about subject of interest and provides insights about the big Mart Problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands for pre-processing of data. Dataset should therefore be explored as much as possible.

1. **computes a summary of statistics :**

Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value etc. It shows five point summary of that big mart Raw data.

A picture containing graphical user interface

Description automatically generated

1. **Checking null Values :**

Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy.

Text

Description automatically generated with low confidence

1. **Data Exploration :**

In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data which shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item Visibility is zero which is not actually practically possible. Establishment year of Outlet varies from 1985 to 2009. These values may not be appropriate in this form. So, we need to convert them into how old a particular outlet is. There are 1531 unique products, as well as 10 unique outlets, present in the dataset. The attribute Item type contains 16 unique values. Where as two types of Item Fat Content are there but some of them are misspelled as regular instead of ’Regular’ and low fat, LF instead of Low Fat. It was found that the response variable i.e. Item Outlet Sales was positively skewed. So, to remove the skewness of response variable a log operation was performed on Item Outlet Sales

1. **Exploratory data analysis (EDA) :**

We have made our hypotheses and now we are ready to do some data exploration and come up with some inference. The goal for the EDA is to get some insight and if any irregularities are found we will correct that in the next section, **Data Pre-Processing.** This shows that our Item Identifier has some duplicate values. since a product can exist in more than one store it is expected for this repetition.

Graphical user interface, text

Description automatically generated

## **Univariate Analysis :**

In Univariate analysis we will explore each variable in a dataset.

1. **Distribution of the target variable: Item\_Outlet\_Sales**

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

We can see that our target variable is skewed towards the right. Therefore, we must normalize it.

1. **Numerical Predictors**

Now we will consider our dependent variables. First of all, we will check for the numerical variables in our dataset:

Graphical user interface, text, application

Description automatically generated

We can see that out of 12 we have only 5 numeric variables.

Graphical user interface

Description automatically generated

A picture containing table

Description automatically generated

**From the above result, we can see that Item\_MRP have the most positive correlation and the Item\_Visibility have the lowest correlation with our target variable. It is totally different from our initial hypotheses, this variables was expected to have high impact in the sales increase. Nevertheless, since this is not an expected behaviour and we should investigate.**

1. **Categorical Predictors**

**Now let’s do some analysis on categorical variable and look at the variables that contain some insight on the hypotheses that we previously made.**

**C 1. Distribution of the variable Item\_Fat\_Content**

Chart, bar chart

Description automatically generated

For Item\_Fat\_Content there are two possible type “Low Fat” or “Regular”. However, in our data it is written in different manner. We will Correct this.

**C2 Distribution of the variable Item\_Type**

Chart, bar chart

Description automatically generated

forItem\_Type we have 16 different types of unique values and it is high number for categorical variable. Therefore we must try to reduce it.

Chart, bar chart

Description automatically generated

There seems to be less number of stores with size equals to “High”. It will be very interesting to see how this variable relates to our target.

**C3. Distribution of the variable Outlet\_Location\_Type**

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedFrom the above graph we can see that Bigmart is a brand of medium and small size city compare to densely populated area.

**Distribution of the variable Outlet\_Type**

There seems like Supermarket Type2 , Grocery Store and Supermarket Type3 all have low numbers of stores, we can create a single category with all of three, but before doing this we must see their impact on target variable.

1. **Bivariate Analysis :**

Now it time to see the relationship between our target variable and predictors.

**Numerical Variables**

Chart, scatter chart

Description automatically generated

We saw previously that Item\_Weight had a low correlation with our target variable. This plot shows there relation.

**8 a. Item\_Visibility and Item\_Outlet\_Sales analysis**

Chart, scatter chart

Description automatically generated

8 b**.** **Outlet\_Establishment\_Year and Item\_Outlet\_Sales analysis**

Graphical user interface

Description automatically generated

There seems to be no appreciable meaning between the year of store establishment and the sales for the items.

8 c. **Categorical Variables**

**Impact of Item\_Fat\_Content onItem\_Outlet\_Sales**

Graphical user interface, application

Description automatically generated

Low Fat products seem to higher sales than the Regular products.

8 d. **Impact of Outlet\_Identifier on Item\_Outlet\_Sales**

Graphical user interface

Description automatically generated with medium confidence

Out of 10- There are 2 Groceries store, 6 Supermarket Type1, 1Supermarket Type2, and 1 Supermarket Type3. You can see from the below pivot table.

Graphical user interface, text, application, Word

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Most of the stores are of Supermarket Type1 of size High and they do not have best results. whereas Supermarket Type3 (OUT035) is a Medium size store and have best results.

**8.e Impact of Outlet\_Size on Item\_Outlet\_Sales**

Graphical user interface, chart, bar chart

Description automatically generated

**8.f Impact of Outlet\_Type on Item\_Outlet\_Sales**

Graphical user interface, chart, bar chart

Description automatically generated

It could be a good idea to create a new feature that shows the sales ratio according to the store size.

Graphical user interface

Description automatically generated

This shows that our hypotheses was totaly different from the result that we got from the above plot. Tier 2 cities have the higher sales than the Tier 1 and Tier 2.

Graphical user interface, text, application

Description automatically generated

1. **Data Cleaning:**

It was observed from the previous section that the attributes Outlet Size and Item Weight has null values. In our work in case of Outlet Size missing value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute. The missing attributes are numerical where the replacement by mean and mode diminishes the correlation among imputed attributes. For our model we are assuming that there is no relationship between the measured attribute and imputed attribute

Chart, box and whisker chart

Description automatically generated

1. **Applying Regression Models:**

**Scikit-Learn can be used to track machine-learning system. Algorithms employed for predicting sales for this dataset are discussed as follows:**

1. **Linear Regression Algorithm**

Regression can be termed as a parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. This technique is said to be parametric as different assumptions are made on basis of data set.

A picture containing text

Description automatically generated

β1 represents the slope.

• β1 represents the estimated change in the average value of y as a result of a one unit change in x.

• β0 represents the estimated average value of y when the value of x is zero.

The values are labeled as shown

Chart

Description automatically generated

Linear regression Model we have applied in our big Mart dataset with following outputs.

A picture containing text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text, letter

Description automatically generated

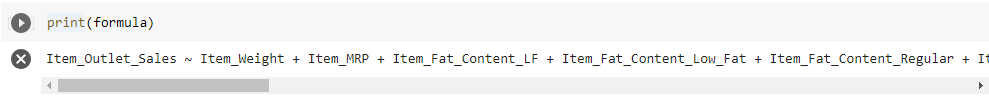
Table

Description automatically generated

Text

Description automatically generated with medium confidence





Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

**B. RIDGE REGRESSION**

It is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

Text

Description automatically generated

**C. LASSO**

Lasso regression is **a type of linear regression that uses shrinkage**. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models

Text

Description automatically generated

# D. Elastic Net

# Elastic net is a popular type of regularized linear regression that combines two popular penalties, specifically the L1 and L2 penalty functions. ... Elastic Net is an extension of linear regression that adds regularization penalties to the loss function during training.

Text

Description automatically generated

1. **Stochastic gradient**

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression.

Graphical user interface, text, application

Description automatically generated

1. **Support Vector Regression (SVR)**

Support Vector Regression is a supervised learning algorithm that is **used to predict discrete values**. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

A screenshot of a computer

Description automatically generated with medium confidence

1. **Decision Tree**

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label.

Graphical user interface, text, application

Description automatically generated

1. **Random forest**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Graphical user interface, text

Description automatically generated

1. **Bagging Regression**

A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

Graphical user interface, text, application

Description automatically generated

1. **Adaptive Boosting**

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

Graphical user interface, text, application

Description automatically generated

1. **Gradient Boosting**

Gradient boosting is a type of machine learning boosting. It relies on **the intuition** that the best possible next model, when combined with previous models, minimizes the overall prediction error. ... If a small change in the prediction for a case causes no change in error, then next target outcome of the case is zero.

Text

Description automatically generated

**11.Metrics for Data Modelling**

The coefficient of determination R2 (R-squared) is a statistic that measures the goodness of a model’s fit i.e. how well the real data points are approximated by the predictions of regression. Higher values of R2 suggest higher model accomplishments in terms of prediction along with accuracy, and the value 1 of R2 is indicative of regression predictions perfectly fitting the real data points. For further better results, the use of adjusted R2 measures works wonders. Taking logarithmic values of the target column in the dataset proves to be significant in the prediction process. So, it can be said that on taking adjustments of columns used in prediction, better results can be deduced. One way of incorporating adjustment could also have included taking square root of the column. It also provides better visualization of the dataset and target variable as the square root of target variable is inclined to be a normal distribution.

The error measurement is an important metric in the estimation period. Root mean squared error (RMSE) and Mean Absolute Error (MAE) are generally used for continuous variable’s accuracy measurement. It can be said that the average model prediction error can be expressed in units of the variable of interest by using both MAE and RMSE. MAE is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The square root of the average of squared differences between prediction and actual observation can be termed as RMSE. RMSE is an absolute measure of fit, whereas R2 is a relative measure of fit. RMSE helps in measuring the variable’s

average error and it is also a quadratic scoring rule. Low RMSE values obtained for linear or multiple regression corresponds to better model fitting.

With respect to the results obtained in this work, it can be said that there is no big difference between our train and test sample since the metric RMSE ratio is calculated to be equal to the ratio between train and test sample. The results related to how accurately responses are predicted by our model can be inferred from RMSE as it is a good measure along with measuring precision and other required capabilities. A considerable improvement could be made by further data exploration incorporated with outlier detection and high leverage points.

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text, table

Description automatically generated

**12 Conclusion & Future Scope**

In this paper, basics of machine learning and the associated data processing and modelling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centers at different locations. On implementation, the prediction results show the correlation among different attributes considered and how a particular location of medium size recorded the highest sales, suggesting that other shopping locations should follow similar patterns for improved sales.

Multiple instances parameters and various factors can be used to make this sales prediction more innovative and successful. Accuracy, which plays a key role in prediction-based systems, can be significantly increased as the number of parameters used are increased. Also, a look into how the sub-models work can lead to increase in productivity of system. The project can be further

collaborated in a web-based application or in any device supported with an in-built intelligence by virtue of Internet of Things (IoT), to be more feasible for use. Various stakeholders concerned with sales information can also provide more inputs to help in hypothesis generation and more instances can be taken into consideration such that more precise results that are closer to real world situations are generated.

When combined with effective data mining methods and properties, the traditional means could be seen to make a higher and positive effect on the overall development of corporation’s tasks on the whole. One of the main highlights is more expressive regression outputs, which are more understandable bounded with some of accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stage of regression model-building. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.