STORES

SALES

PREDICTION

PROJECT REPORT

CONTENTS

INTRODUCTION

Due to the quick rise of global malls and on-line shopping, competition among different shopping malls as well as huge marts is becoming more serious and hostile day by day. Every mall or supermarket tries to attract more people by offering personalized and limited-time offers based on the day, so that the volume of sales for each item can be predicted for inventory management, logistics, and transportation, among other things. Machine learning algorithms are quite powerful today, and they provide ways for predicting or forecasting future sales demand for an organization, as well as overcoming the low cost of computer and storage systems.

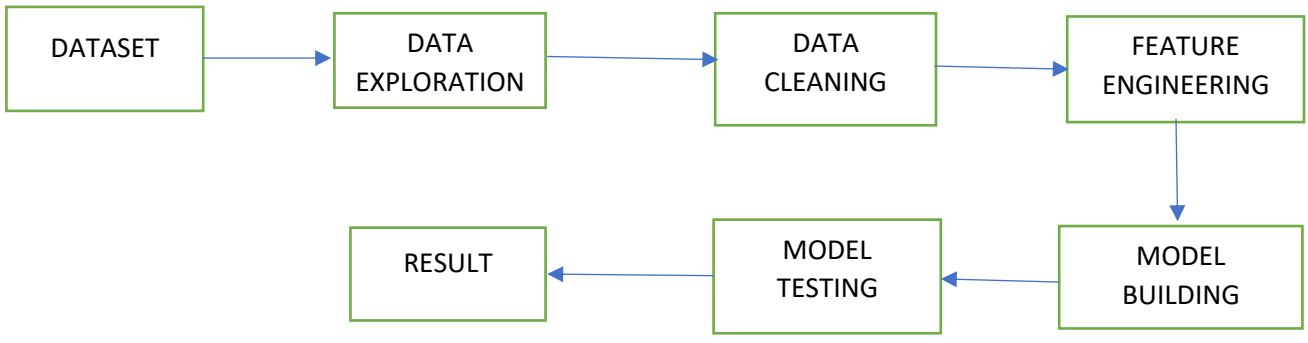
A store sales prediction study can aid in the in-depth analysis of previous scenarios or conditions, and then the inferences about client acquisition, funds inadequacy, and strengths can be implemented before establishing a budget and marketing plans for the following year. To put it another way, sales forecasting is based on historical resources. In this project, we look at the problem of predicting sales or estimating an item's future demand based on historical data in different stores across multiple locations and goods. For predicting or forecasting sales volume, many machine learning methods such as linear regression analysis, random forest, and others are utilized. Because good sales are the lifeblood of every business, sales forecasting is critical in any shopping centre.

PROBLEM STATEMENT

Currently, shopping promenades and big marts keep track of individual item deals data in order to read unborn customer demand and acclimate force operation in a data storehouse these data stores hold a significant quantum of consumer information and particular item details by booby-trapping the data store from the data storehouse, further anomalies and commo patterns can be discovered

DATA DESCRIPTION

The dataset consists of 12 attributes like Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales.

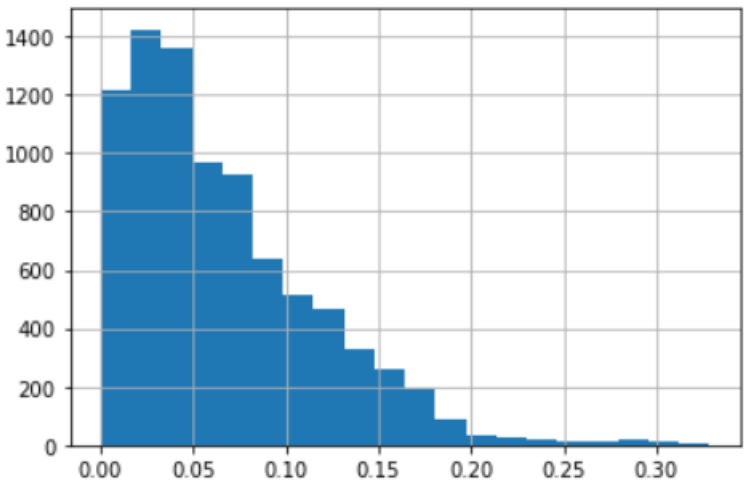


DATA EXPLORATION

The attributes Outlet size and Item weight are missing values, and Item Visibility's minimum value is 0, which is not really conceivable. Outlet's founding year ranges from 1985 to 2009. In this format, certain values might not be acceptable. As a result, we'll need to translate them to the age of a specific outlet. The collection contains 1559 unique products as well as 10 distinct outlets. There are 16 distinct values in the attribute Item type. Whereas there are two sorts of Item Fat Content, some of them are misspelt, such as regular instead of 'Regular' and low fat, LF instead of Low Fat. The response variable, Item Outlet Sales, was found to be positively skewed. A log operation on Item Outlet Sales was used to remove the skewness of response variable.

DATA CLEANING

The attributes Outlet Size and Item Weight have missing values, as seen in the previous section. In our work, we replace missing Outlet Size values with the mode of that attribute, and we replace missing Item Weight values with the mean of that property. The missing attributes are numerical, and replacing them with mean and mode reduces the correlation between the imputed data. We assume that there is no link between the measured and imputed attributes in our model.



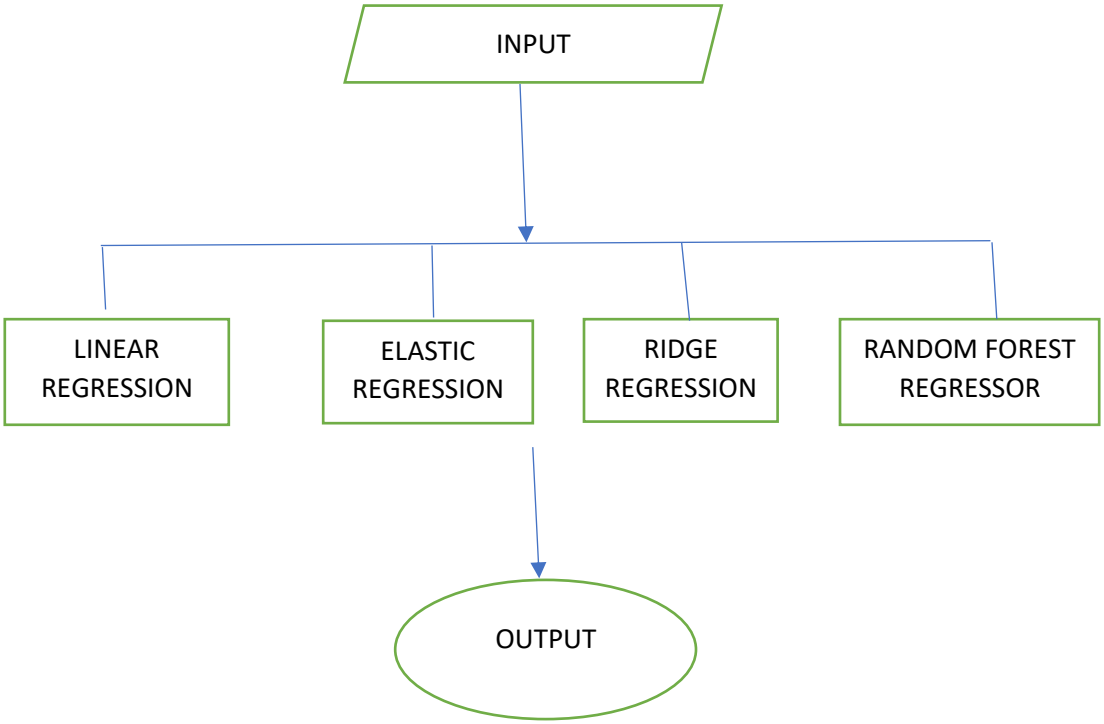
FEATURE ENGINEERING

During the data exploration phase, some idiosyncrasies in the data set were discovered. As a result, this phase is used to resolve all nuances discovered in the dataset and prepare them for the creation of the suitable model. During this step, it was discovered that the Item visibility attribute was set to zero, which makes no sense. As a result, for zero values attribute, the product's mean value item visibility will be used. As a result, all products are more likely to sell. All discrepancies in categorical attributes are handled by converting all categorical attributes to acceptable ones. Non-consumables and fat content property properties are not specified in some circumstances. To avoid this, we add a third Item fat content category: none. The unique ID starts with either DR, FD, or NC, according to the Item Identifier attribute. As a result, we construct a new characteristic called Item Type New, which has three categories: foods, beverages, and non-consumables. Finally, we add a new attribute Year to the dataset to determine how old a certain outlet is.

MODEL BUILDING

The dataset is now ready for the proposed model to be built. The model is then utilized as a predictive model to forecast Big Mart sales. We present a model employing the XgBoost method and compare it to various machine learning techniques such as linear regression, ridge regression, and elastic regression and so on.

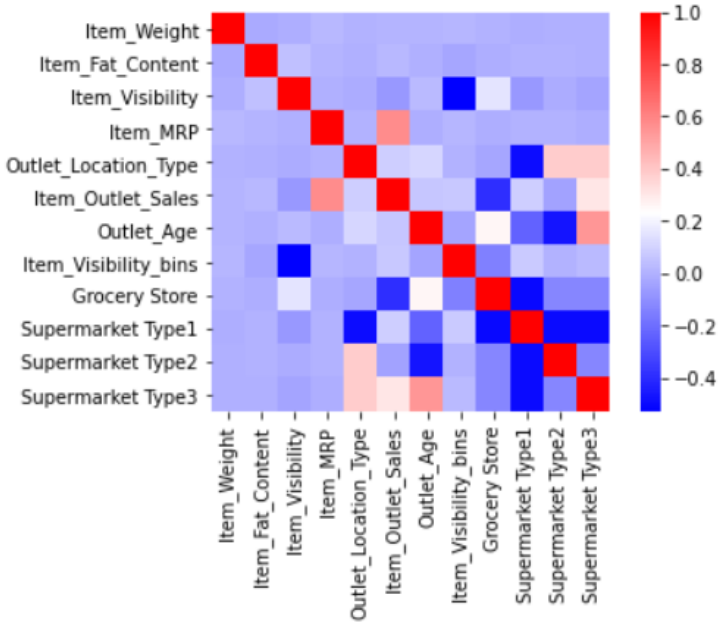
Linear Regression: A model which create a linear relationship between the dependent variable and one or more independent variable.



RESULTS

Every model is first trained with training data before being used to forecast accuracy with test data, and this process is repeated until each subset has been tested once. The smallest sales were created in the smallest places, according to data visualization. However, in some cases, it was discovered that a medium-sized location, even though it was a type-3 (there are three types of supermarkets: type-1, type-2, and type-3) super market, produced the highest sales, and that more locations should be switched to Type 3 Supermarkets to increase Big Mart product sales in a particular outlet.

	Grocery Store	Supermarket Type1	Supermarket Type2	Supermarket Type3
0	0	1	0	0
1	0	0	1	0
2	0	1	0	0
3	1	0	0	0
4	0	1	0	0

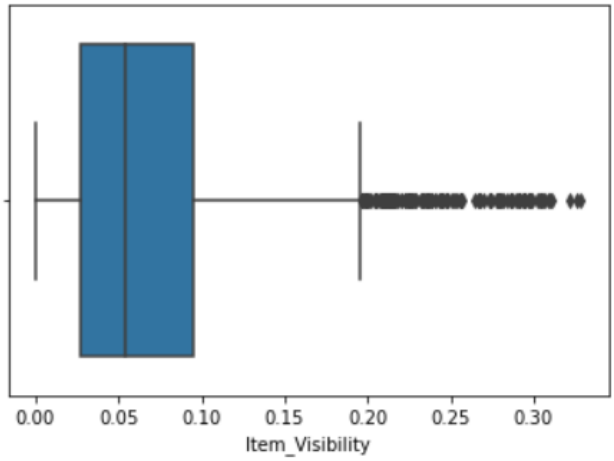


CONCLUSION

In today's digitally connected world, every shopping mall wishes to anticipate customer wants in order to minimize seasonal shortages of sale items. Companies and shopping malls are getting better at anticipating product sales and consumer requests on a daily basis. For precise sales forecasting, extensive research is being conducted at the organization level. Because a company's profit is directly linked to how accurate its sales projections are, Big Marts want a more accurate prediction algorithm so that they don't lose money.

	Item_Outlet_Sales
0	3735.1380
1	443.4228
2	2097.2700
3	732.3800
4	994.7052
...	...
8518	2778.3834
8519	549.2850
8520	1193.1136
8521	1845.5976
8522	765.6700

8379 rows × 1 columns



REFERENCES

- Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., L¨utjen, M., Teucke, M.: A surveyon retail sales forecasting and prediction in fashion markets. Systems Science &Control Engineering 3(1), 154–161 (2015)

- Ni, Y., Fan, F.: A two-stage dynamic sales forecasting model for the fashion retail.Expert Systems with Applications 38(3), 1529–1536 (2011)

- Shrivas, T.: Big mart dataset@ONLINE (Jun 2013), https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

- Sekban, Judi. "Applying machine learning algorithms in sales prediction." (2019).