BIG MART SALES ANALYSIS

# TABLE OF CONTENTS

1. INTRODUCTION
2. ABSTRACT
3. BACKGROUND
4. DATA SET DESCRIPTION
5. DATASET SOURCE
6. SAMPLE DATAS
7. DATASET CHARCTERISATION
8. LETERATURE SURVEY
9. PROPOSED ALGORITHM
10. RESULT
11. CONCLUSION
12. SUMMARY
13. REFERENCES

# INTRODUCTION

A Big mart is a shopping mall which sells variety of all household, eatables, electronic devices, Garments, Groceries at a large scale. But the sales of a product may vary season to season. For instance, Large scale of Air conditioners will be bought by the customers during summer and less in winter. When the sales of products vary, the employees of big mart may not know what the sales forecast is and how much production is needed in the stock. In this case, sales forecasting plays an important role to predict the sales of each and every product by the help of cumulative sales report. To predict future sales, an algorithm is required to predict the sales and in order to get accurate results. Decision trees are basically predictive machine learning models. Decision trees models helps to predict a class for the case after training pruning and testing is over.

# ABSTRACT

Big mart sales prediction is about predicting future sales using the cumulative sales reports. The datasets from kaggle repository where the datasets about 1559 products and 10 outlets/stores were taken. We are using Pandas for handling data and numpy for handling numerical operations in arrays. The algorithm which is going to be used is Machine Learning Algorithms such as decsion tree regression,linear regression,random forest tree regressor and Extra tree regressor. Regression is used to predict a range of numerical values, given a particular dataset. Decision tree is linearized into decision rules where the outcome is the contents of the leaf node and the conditions along the path form the conjunction in the if clause. The aim is to build a predictive model and find the sales of each product at a particular store. Using this model, big marts will try to understand the properties of the products and stores which play a key role in increasing sales, where to improve the marketing or to stop the selling of the product.

**Keywords** – Sales forecast, Decision tree Regression, Pandas, Numpy, datasets, Cumulative sales.

**BACKGROUND**

We are using pandas for handing data and numpy for handling numerical operations in arrays

**PANDAS**

Python has long been great for data munging and preparation, but less so for data analysis and modeling. pandas helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R.

Combined with the excellent IPython toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity, and the ability to collaborate. pandas does not implement significant modeling functionality outside of linear and panel regression; for this, look to stats models and scikit learn. More work is still needed to make Python a first class statistical modeling environment, but we are well on our way toward that goal.

**NUMPY**

NumPy is the fundamental package for scientific computing with Python. It contains among other things a powerful N-dimensional array object Sophisticated (broadcasting) functions

Tools for integrating C/C++ and Fortran code

Useful linear algebra, Fourier transform, and random number capabilities . Besides its obvious scientific uses, NumPy can also be used as an efficient multidimensional container of generic data. Arbitrary data types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

**DATA SET DESCRIPTION**

NO OF COLUMNS (12) NO OF ROWS (8524)

DATA SET CHARACTERISTICS (MULTI VARIATE) ATTRIBUTE CHARACTERISTICS (INTEGER, REAL) ASSOCIATED TASKS (CLASSIFICATION, CLUSTERING) NUMBER OF ATTRIBUTES (12)

ITEM\_IDENTIFIER ITEM\_WEIGHT ITEM\_FAT\_CONTENT ITEM\_VISIBILITY ITEM\_TYPE ITEM\_MRP OUTLET\_IDENTIFIER

OUTLET\_ESTABLISHMENT\_YEAR OUTLET\_SIZE OUTLET\_LOCATION\_TYPE OUTLET\_TYPE ITEM\_OUTLET\_SALES

**SAMPLE DATA**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Item\_ Weig ht | Item\_ Fat\_ Conte nt | Item\_ Visibi lity | Item  \_ Type | Item  \_ MRP | Outlet\_ Identifi er | Outlet  \_Establi sh ment  \_Year | Outle t\_ Size | Outl et  \_Typ e | Outlet  \_Locati on  \_Type | Item\_Outlet\_S ales |
| 9.3 | 1 | 0.016  05 | 1 | 249.8  1 | 49 | 1999 | 2 | 1 | 1 | 3735.138 |
| 5.92 | 2 | 0.019  28 | 2 | 48.26  9 | 18 | 2009 | 2 | 2 | 3 | 443.4228 |
| 17.5 | 1 | 0.016  76 | 3 | 141.6  2 | 49 | 1999 | 2 | 1 | 1 | 2097.27 |
| 19.2 | 2 | 0 | 4 | 182.1 | 10 | 1998 |  | 4 | 3 | 732.38 |
| 8.93 | 1 | 0 | 5 | 53.86  1 | 13 | 1987 | 1 | 1 | 3 | 994.7052 |
| 10.39  5 | 2 | 0 | 6 | 51.40  1 | 18 | 2009 | 2 | 2 | 3 | 556.6088 |
| 13.65 | 2 | 0.012  74 | 7 | 57.65  9 | 13 | 1987 | 1 | 1 | 3 | 343.5528 |
|  | 1 | 0.127  47 | 7 | 107.7  6 | 27 | 1985 | 2 | 3 | 3 | 4022.764 |
| 16.2 | 2 | 0.016  69 | 8 | 96.97  3 | 45 | 2002 |  | 1 | 2 | 1076.599 |
| 19.2 | 2 | 0.094  45 | 8 | 187.8  2 | 17 | 2007 |  | 1 | 2 | 4710.535 |
| 11.8 | 1 | 0 | 4 | 45.54 | 49 | 1999 | 2 | 1 | 1 | 1516.027 |
| 18.5 | 2 | 0.045  46 | 1 | 144.1  1 | 46 | 1997 | 3 | 1 | 1 | 2187.153 |
| 15.1 | 2 | 0.100  01 | 4 | 145.4  8 | 49 | 1999 | 2 | 1 | 1 | 1589.265 |
| 17.6 | 2 | 0.047  26 | 7 | 119.6  8 | 46 | 1997 | 3 | 1 | 1 | 2145.208 |
| 16.35 | 1 | 0.068  02 | 4 | 196.4  4 | 13 | 1987 | 1 | 1 | 3 | 1977.426 |
| 9 | 2 | 0.069  09 | 9 | 56.36  1 | 46 | 1997 | 3 | 1 | 1 | 1547.319 |
| 11.8 | 1 | 0.008  6 | 10 | 115.3  5 | 18 | 2009 | 2 | 2 | 3 | 1621.889 |

# DATA SOURCE

The datasets were downloaded from Kaggle repository.

Kaggle is an online community of data scientists and machine learners. It allows users to find and publish data sets, explore and build models in a web-based data science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

REFERENCE LINK:

https://[www.kaggle.com/devashih0507/big-mart-sales-prediction](http://www.kaggle.com/devashih0507/big-mart-sales-prediction)

**CHARACTERIZATION:**

|  |  |
| --- | --- |
| Variable | Description |
| Item\_identifier | Unique produce id |
| Item\_weight | Weight of product |
| Item\_fat\_content | Whether the product is low fat or not |
| Item\_visibility | The % of total display area of all products in a store allocated to the particular product |
| Item\_type | The category to which the product belongs |
| Item\_MRP | Maximum retail price of the product |
| Outlet\_identifier | Unique store id |
| Outlet\_establishment\_year | The year in which store was established |
| Outlet\_size | The size of the store in terms of ground area covered |
| Outlet\_location\_type | The type of city in which the store is located. |
| Outlet\_type | Whether the outlet is just a grocery store or some sort of super market. |
| Item\_outlet\_sales | Sales of the product in the particular store. This is the outcome variable to be predicted. |

# RESEARCH PAPERS:

1.Predicting sales in super market(2018)

2.A Two-Level Statistical Model for Big Mart Sales Prediction(2019) 3.Sales Prediction System using Machine Learning (2019)

1. Machine-Learning Models for Sales Time Series Forecasting(2019)
2. Walmart’s Sales Data Analysis- A Big Data Analytics Perspective(2017)
3. Sales Forecasting for Retail Chains(2018)
4. A Comparative Study of Big Mart Sales Prediction
5. An Ensemble Based Predictive Modeling in Forecasting Sales of Big Mart.
6. Outlet Sales Analysis using R and Various Machine Learning Algorithms
7. Secure Web Based Sales Configurat ions in a Trading Industry (2016)
8. An Ensemble Based Predictive Modeling in Forecasting Sales of Big Mart (2016)

# LITERATURE REVIEW:

**PAPER 1:**

There are two noteworthy types of forecasting. Macro Forecasting and Micro Forecasting. Macro Forecasting is unstable with foreseeing commercial centres in entirety. This will be identified with close the present level of Market Plea and assessing the results of market request in the up and coming ages. Micro Forecasting is concerned with entire segment deals figures. This will be tied in with moulding a thing's business sector parcel specifically industry and thinking about the outcomes of piece of the overall industry later on period. A gauge is an evaluation of an episode which will occur in future. The forecast esteem isn't a deterministic measure. Since it is just a gauge in light of the past information identified with a specific occasion, legitimate care must be given in evaluating it. All the practical chiefs in

any association will construct their choices with respect to the figure esteem. Along these lines, it is fundamental data for the affiliation.

Because of these reasons, appropriate upkeep ought to be executed while evaluating estimate models.

# PAPER 2:

In this paper, the ensemble of data mining predictive techniques via stacking is considered a two-level statistical approach. It is named as two-level because stacking is performed on two layers in which bottom layer consists of one or more than one learning algorithms and top layer consists of one learning algorithm. Stacking is also known as Stacked Generalization. It basically involves the training of the learning algorithm present in the top layer to combine the predictions made by the algorithms present in the

bottom layer. In the first step, all the learning algorithms are trained using the big mart dataset and in the second step, a combiner algorithm is trained using all the predictions made by the bottom layer algorithms to get a final prediction. Stacking performs better than any single model because a stacking involves more information for prediction. Each algorithm at lower level operates differently and makes individual predictions. All aspects of the dataset are viewed using various learning algorithms and then the final prediction is made.

# PAPER 3:

Forecasting sales become hard manually when the number of factors increases. Demand prediction is also closely related to Sales revenue. If sellers store much more product than the demand then this may lead to surplus. On the other hand, storing less product in order to save inventory costs when the product has a high demand will Thus, Sales Prediction helps the companies to store products according to expected sales for the region and outlet type.Thus, providing companies with the predicted sales for products and different outlet locations helps companies to formulate a proper business model which helps them to organize.

# PAPER 4:

In this paper[4], we study the usage of machine-learning models for sales predictive analytics. The main goal of this paper is to consider main approaches and case studies of using machine learning for sales forecasting. The effect of machine-learning generalization has been considered. This effect can be used to make sales predictions when there is a small amount of historical data for speciﬁc sales time series in the case when a new product or store is launched. A stacking approach for building regression ensemble of single models has been studied. The results show that using stacking

techniques, we can improve the performance of predictive models for sales time series forecasting.

# PAPER 5 :

To create effective promotions and offers to meet its sales and marketing goals, otherwise they will forgo the major opportunities that the current market offers. Big Data application enables these retail organizations to use prior year’s data to better forecast and predict the coming year’s sales. It also enables retailers with valuable and analytical insights, especially determining customers with desired products at desired time in a particular store at different geographical locations. In this paper [5], we analyzed the data sets of world’s largest retailers, Walmart Store to determine the business drivers and predict which departments are affected by the different scenarios (such as temperature, fuel price and holidays) and their impact on sales at stores of different locations.

# PAPER 6:

In paper [6], we have performed sales forecasting for stores using different data mining techniques. The task involved predicting the sales on any given day at any store. In order to familiarize ourselves with the task we have studied previous work in the domain including Time Series Algorithm as well as a Spatial approach. A lot of analysis was performed on the data to identify patterns and outliers which would boost or impede the prediction algorithm.

The features used ranged from store information to customer information as well as socio- geographical information. Data Mining methods like Linear Regression, Random Forest Regression and XGBoost were implemented and the results compared. XGBoost which is an improved gradient boosting algorithm was observed to perform the best at prediction. With efficiency being the way forward in most industries today, we aim to expand our solution to

help stores improve productivity and increase revenue by taking advantage of Data Analysis.

# PAPER 7:

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service, etc. Present machine learning algorithm are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organization, which also helps in overcoming the cheap availability of computing and storage systems.

# PAPER 8:

The statistical learning of the ensemble modeling to predict the sales of BigMart can be beneficially adopted for the wholesale and retail vendor joints in India. It helps in understanding the factors that influence the sales of similar products manner. The machine learning techniques adopted aims at reducing the variability to the maximum in predicting the sales. In every technique that has been adopted, all the related issues has been addressed. Accuracy of the prediction is determined by fitting the development model on the valuation data.

# PAPER 9:

As we know that in today’s era data analysis is so important to everyone to make better decisions in their field. Analyzing the big data and extracting knowledge full information from the data is little bit tough. so, for mining of complex datasets we need a powerful and effective data mining tool to extract the information and take better decisions in future. We are using R here which is an open source free data mining tool and efficient too. R has several inbuilt packages which provides us efficiency like-ggplot2, VIM etc. R is an open-source data analysis environment and programming language .The process of converting data into knowledge, insight and understanding is Data analysis, which is a critical part of statistics. For the effective processing and analysis of big data, it allows users to conduct a number of tasks that are essential. R consists of numerous ready-to-use statistical modeling algorithms and machine learning which allow users to create reproducible research and develop data products.

# PAPER 10:

The Apriori algorithm, known to be one of the best algorithms for finding frequent item sets in large transactional databases, is the subject of research in this paper [18]. While carrying out of this process, some sensitive information face disclosure in front of the unauthorized users which should be taken care of and this is carried out by the process of privacy preservation. It accumulates a huge amount of sales data on a daily basis, there is a constant need for analyzing data to determine frequently purchased items by a customer, over a period of time.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | S.NO | Paper title | Algortihm used | Data set beign used | Performance and measure | Gap identified | Paper scope |  |  |
| 1. | Predicting Sales in  Supermarkets (2018) | ANN  algorithm | BigMart Sales Data | Root  mean square error. | Most companies like to estimate the upcoming trades. A superior forecasting can avoid them from overestimating Or underestimating the future | The paper aim to put on different machine learning techniques to construct and adjust a sales forecasting model and perform estimation on sales data to come across this requirement. We also intend to provide a simple to use result with various visualization tools for the easiness of users. |
| 2. | A Two-Level Statistical Model for Big Mart Sales Prediction  (2019) | Machine learning algorithm | Big mart sales | Mean absolute error | The aim is to build a predictive model and find out the sales of each product at a particular store. | An effort has been made to predict sales of the product from a particular outlet accurately by using a  two-level statistical model that reduces the mean absolute error value up to acceptable. |
| 3. | Sales Prediction System using Machine Learning (2019) | Linear Regression and XG boost Regressor | BigMart Sales Data | Root  mean square error. | Supply and demand are two fundamenta l concepts of sellers and customers.  Predicting demand are accurate is critical for organization s in order to | In this paper, we propose a new approach for demand prediction for Big Mart companies. The experimental results show that the XGBoost regressor gives pretty accurate sales results. |
|  | | | | | | | |
|  | | | | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | be able to formulate plans. |  |  |
| 4 | Machine-Learning Models for Sales Time  Series Forecasting (2019) | Machine learning algorithm | BigMart Sales Data | Root mean square error. | Less performanc e of predictive models for sales time series forecasting. | The main goal of this paper is to consider main approaches and case studies of using machine learning for sales forecasting. The effect of machine- learning generalization has been considered |
| 5 | Walmart’s Sales  Data Analysis- A Big Data Analytics Perspective(2017) | Technique  -Hadoop Distributed File Systems; Apache Spark; MapReduce | BigMart  Sales Data | Root mean  square error | Many times it is hard for the retailers to comprehend the market condition since their retail stores are at various geographical locations. | In this paper, we analysed the data sets of world’s largest retailers, Walmart Store to determine the business drivers and predict which departments are affected by the different scenarios and their impact on sales at stores of different locations. |
| 6 | Sales Forecasting for Retail Chains | Gradient Boosting algorithm, XGBoost algorithm | BigMart Sales Data | Root mean square error | Findings not only reveal that the XGBoost algorithm  outperforms the traditional modeling approaches with regard to prediction accuracy, but it also uncovers  new | This paper presents a use case of data mining for sales forecasting in retail demand and sales prediction |
|  | | | | | | | | |

1. A

Comparative Study of Big Mart Sales Prediction

Xgboost algorithm and GBM algorithm

BigMart dataset

Root Mean Square Error(RMSE).

Most

of these models are linear and are not able to deal with the asymmetric behavior

in most real-world sales data. Some of the challenging factors like lack of historical data, consumer-oriented markets face uncertain demands, and short life cycles of prediction methods results in inaccurate forecast.

As good sales are the life of every organization so the forecasting of sales plays an important role in any

shopping complex. Always a better prediction is helpful, to develop as well as to enhance the strategies of business about the marketplace which is also helpful

2 Gopal Behera and Neeta Nain

to improve the knowledge of marketplace.

1. An

Ensemble Based Predictive Modeling in Forecasting Sales of Big Mart.

1. Outlet Sales Analysis using R and Various Machine Learning Algorithms

random forest, support vector machine and neural network

Random forest, ANN, SVM

machine learning, K-means,

Naive Bayes algorithm.

Test dataset

big mart sales dataset.

Root Mean Square Error(RMSE).

Root Mean Square Error(RMSE).

The

regression model is built with transformed variables. It is made

obvious by plotting the residuals against the variables.

From the

model summary it

The main motive of this paper is to show you how to tackle or deal with such giant dataset which has missing values as well as regularities.

The machine learning techniques

adopted aims at reducing the variability to the maximum in

predicting the sales.

In every technique that has been adopted, all the related issues

has been addressed. Accuracy of the prediction is determined

After importation we have to explore and visualize the dataset to find out missing values and irregularities in it. If , the irregularities and missing values are present in the dataset then we have to clean it and impute the data at the place of missing value , after that we can apply several data mining algorithms for prediction.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10. | Secure Web Based Sales Configurat  ions in a Trading | Apriori algorithm | BigMart Sales Data | Root mean square error. | Today every industry, be it Retail or Banking or Trading, has to | This research paper proposes Frequent pattern Tree structure in data mining used to satisfy minimum support and less number of database scans. It also explains the security by providing verification secret code during login by using random password (Session Pair based authentication) that leads to a secure application. |  |  |
|  | Industry |  |  |  | deal with huge |
|  | (2016) |  |  |  | sets of data. The |
|  |  |  |  |  | data has to be |
|  |  |  |  |  | analyzed and |
|  |  |  |  |  | useful |
|  |  |  |  |  | information has |
|  |  |  |  |  | to be extracted |
|  |  |  |  |  | to understand |
|  |  |  |  |  | customer’s |
|  |  |  |  |  | interests and |
|  |  |  |  |  | meet their |
|  |  |  |  |  | demands in a |
|  |  |  |  |  | better way. |
|  | | | | | | | |
|  | | | | | | | | | |

# PREPROCESSING

Pre- processing is performed on the cumulative sales data sets in order to remove missing values, outliers and noisy data’s.In the current dataset, many missing values were found in many tuples under the attributes.We have applied mean formulae to calculate the missing values.

The mean of the whole column was calculated to provide data to the missing cell. This process was applied iteratively until no missing cells were left under attributes.

Scaling is performed to the values for fitting and transforming the dataset. In this case, the data’s under an attribute are taken. Mean and standard deviation is performed. The data’s are scaled and transformed until we get mean value as 0 and standard deviation value as 1.

# PROPOSED ALGORITHM MACHINE LEARNING METHODS DECISION TREE REGRESSION

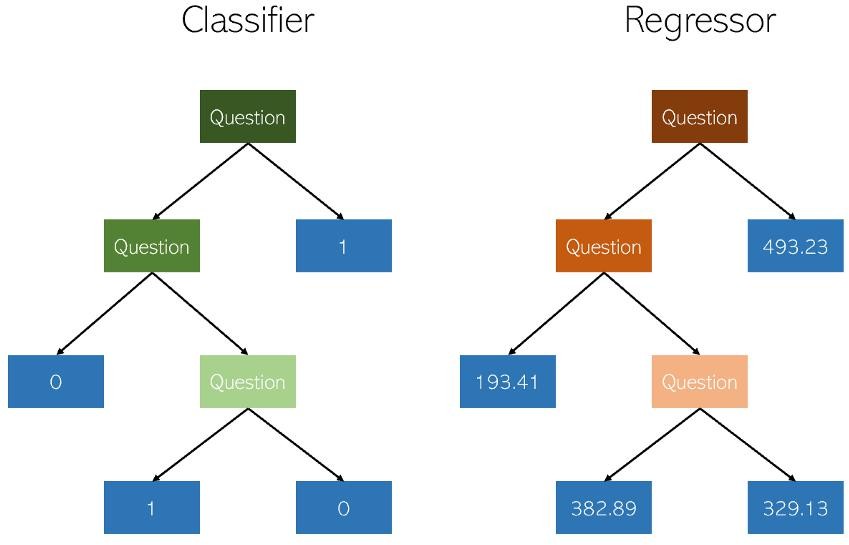
Decision trees are basically predictive machine learning models. Decision trees models helps to predict a class for the case after training pruning and testing is over. It is mainly of two types:

* 1. CLASSIFICATION TREE and 2) REGRESSION TREES.

In case data is continuous type with associated classes also numerical type. For example if target is to predict sales forecast of big mart or price of a house or setting of an apparatus mostly Regression type DECISION TREES are preferred.

The main difference between a regression tree and a classification tree is the how you measure the "badness" of a node. There are various ways to do it for both regression and classification trees. For regression trees, sum of squared error or median absolute deviation or some other function is used.

In a regression tree the idea is this: since the target variable does not have classes, we fit a regression model to the target variable using each of the independent variables. Then for each independent variable, the data is split at several split points. At each split point, the "error" between the predicted value and the actual values is squared to get a "Sum of Squared Errors ". The split point errors across the variables are compared and the variable/point yielding the lowest SSE is chosen as the root node/split point. This process is recursively continued.



# LINEAR REGRESSION

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other, but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables, then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two

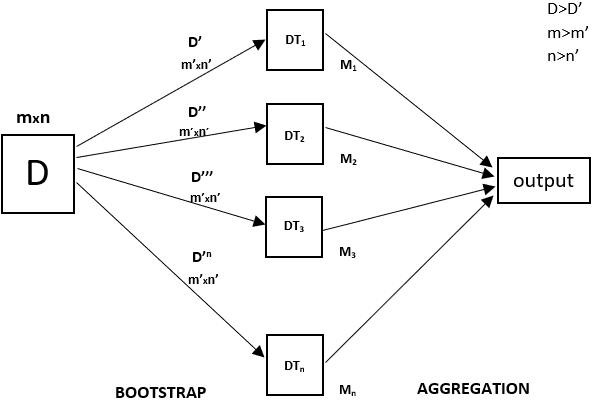
variables.

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

# RANDOM FOREST REGRESSION

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn’t depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is Aggregation.

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.



# EXTRA REGRESSION

Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de- correlated decision trees collected in a “forest” to output it’s classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, Each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index). This random sample of features leads to the creation of multiple de-correlated decision trees.

To perform feature selection using the above forest structure, during the construction of the forest, for each feature, the normalized total reduction in the mathematical criteria used in the decision of feature of split (Gini Index if the Gini Index is used in the construction of the forest) is computed. This value is called the Gini Importance of the feature. To perform feature selection, each feature is ordered in descending order according to the Gini Importance of each feature and the user selects the top k features according to his/her choice.

# Performance Metrics:

MEAN SQUARE ERROR :

The mean square error (MSE) is just like the MAE, but squares the difference before summing them all instead of using the absolute value. The effect of the square term in the MSE equation is most apparent with the presence of outliers in our data. While each residual in MAE contributes proportionally to the total error, the error grows quadratically in MSE.

This ultimately means that outliers in our data will contribute to much higher total error in the MSE than they would the MAE. Similarly, our model will be penalized more for making predictions that differ greatly from the corresponding actual value.

K-FOLD CROSS-VALIDATION

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross- validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

# RESULT:

|  |  |  |
| --- | --- | --- |
| **ALGORTIHM** | **MEAN SQUARE ERROR** | **CROSS VALIDATION SCORE** |
| Linear Regressor | MSE: 0.2880065032501795 | CV Score: 0.778987758788712 |
| Desicion Tree Regressor | MSE: 9.251544668064852e-11 | CV Score: 0.902351523453423 |
| Random Forest Regress or | MSE: 0.04215373973996043  6 | CV Score: 0.695998432444653  6 |
| ExtraTreesRegressor | MSE: 8.744068859353917e-13 | CV Score: 0.731855289221172  7 |

**CONCLUCSION**

We have analyzed datasets of big mart sales prediction and performed to sales prediction using various techniques such as Desicion Tree Regressor, Linear Regressor, Random Forest Tree Regressor and Extra Regressor. We used Jupyter tool through Anaconda Navigator for processing the techniques. Decision Tree Regression proved the best model to predict the future sales with the accuracy rate of 90%. Training the model was easier than any other models. It proved to be the best model in forecasting sales of Big Mart. This indirectly helps to gain more profit and have a scheduled products in stock.

# SUMMARY

A Big mart is a shopping mall which sells variety of all household, eatables, electronic devices, Garments, Groceries at a large scale. But the sales of a product may vary season to season. In this case, sales forecasting plays an important role to predict the sales of each and every product by the help of cumulative sales report. The algorithm which was used in this thesis is Decision Tree regression. Regression is used to predict a range of numerical values, given a particular dataset. The aim was to build a predictive model and find the sales of each product at a particular store. Using this model, big marts will try to understand the properties of the products and stores which play a key role in increasing sales, where to improve the marketing or to stop the selling of the product.

# REFERENCE LINK:

1. [https://www.researchgate.net/publication/336530068\_A\_Comparative\_Study](https://www.researchgate.net/publication/336530068_A_Comparative_Study_of_Big_Mart_Sales_Prediction)

\_o f\_Big\_Mart\_Sales\_Prediction

1. [https://www.worldwidejournals.com/international-journal-of-scientific-](https://www.worldwidejournals.com/international-journal-of-scientific-research-(IJSR)/recent_issues_pdf/2016/May/May_2016_1492767915__225.pdf) research (IJSR)/recent\_issues\_pdf/2016/May/May\_2016\_1492767915 225.pdf

3. <https://acadpubl.eu/hub/2018-119-16/2/386.pdf>

1. <https://ieeexplore.ieee.org/document/5541170>
2. [https://www.worldwidejournals.com/international-journal-of-scientific-](https://www.worldwidejournals.com/international-journal-of-scientific-research-(IJSR)/article/an-ensemble-based-predictive-modeling-in-forecasting-sales-of-big-mart/ODE0NQ%3D%3D/?is=1) research-(IJSR)/article/an-ensemble-based-predictive-modeling-in- forecasting-sales-of- big-mart/ODE0NQ==/?is=1
3. <https://irjet.net/archives/V5/i3/IRJET-V5I3226.pdf>
4. [https://www.researchgate.net/publication/332583719\_A\_Two-](https://www.researchgate.net/publication/332583719_A_Two-Level_Statistical_Model_for_Big_Mart_Sales_Prediction) Level\_Statistical\_M odel\_for\_Big\_Mart\_Sales\_Prediction
5. [https://www.researchgate.net/publication/319990412\_Application\_of\_Art](https://www.researchgate.net/publication/319990412_Application_of_Artificial_Neural_Networks_Using_Bayesian_Training_Rule_in_Sales_Forecasting_for_Furniture_Industry) ificial\_Neural\_Networks\_Using\_Bayesian\_Training\_Rule\_in\_Sales\_Forecast ing\_for\_Furnitur e\_Industry
6. [https://www.researchgate.net/publication/330484523\_Machine-](https://www.researchgate.net/publication/330484523_Machine-Learning_Models_for_Sales_Time_Series_Forecasting) Learning\_Models\_ for\_Sales\_Time\_Series\_Forecasting
7. [https://www.researchgate.net/publication/331606166\_Intelligent\_Sales\_](https://www.researchgate.net/publication/331606166_Intelligent_Sales_Prediction_Using_Machine_Learning_Techniques) Prediction\_ Using\_Machine\_Learning\_Techniques
8. [https://www.semanticscholar.org/paper/Sales-Forecasting-for-Retail-](https://www.semanticscholar.org/paper/Sales-Forecasting-for-Retail-Chains-Jain-Menon/76a244f4da1d29170a9f91d381a5e12dc7ad2c0f) Chains-Jain- Menon/76a244f4da1d29170a9f91d381a5e12dc7ad2c0f
9. [https://www.researchgate.net/publication/328246040\_Walmart's\_Sales\_](https://www.researchgate.net/publication/328246040_Walmart%27s_Sales_Data_Analysis_-_A_Big_Data_Analytics_Perspective) Data\_Anal ysis\_-\_A\_Big\_Data\_Analytics\_Perspective