

**HO CHI MINH UNIVERSITY OF TECHNOLOGY AND EDUCATION FACULTY OF INTERNATIONAL EDUCATION**

🙠🙟🕮🙝🙢



**MACHINE LEARNING**

**FINAL PROJECT**

**(BIG MART PREDICTION)**

|  |  |  |
| --- | --- | --- |
| **Lecturer** : | ThS. Quach Dinh Hoang |  |
| **Student:** | Ho Thanh Duy | 21110759 |
|  | Nguyen Phuoc Truong | 21110 |
|  | Dang Trung Phuong | 21110 |
| **Class code** : | MALE431085E\_23\_1\_02FIE | |
| **Course year:** | 2022-2023 |  |

**LECTURER COMMENT**

Mark: ……………………

**SIGNATURE OF LECTURER**

**ACKNOWLEDGEMENT**

Firstly, We are really grateful and wish our profound indebtedness to ThS. Quach Dinh Hoang, University of Technology and Education. Deep Knowledge & keen interest of my supervisor in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to my teacher, for his kind help to finish this project. We would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, We might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient helpand facilitated my undertaking.

**CHAPTER 1: INTRODUCTION**

1.1. INTRODUCTION

The intense competition that exists between small and large malls on a daily basis is a result of the explosive growth of online shopping and international supermarkets. Every mall or mart strives to offer one-off, temporary donations or benefits to draw in an increasing number of patrons on a daily basis. Examples of these include the sales price of all items, which is typically estimated to be managed through various channels like corporate asset management, logistics, and transportation services, among others. These days, sophisticated machine learning algorithms that offer methods for forecasting or projecting long-term demand for a business's sales also assist in overcoming financial constraints and computer programs.

This report primarily addresses the topic of identifying a large-mart sale or forecasting a future-need item for a customer in a few different supermarket locations and goods that back up the prior record. To predict sales volume, a variety of machine learning algorithms are employed, including *random forest* and *linear regression*. Since effective marketing is, as we all know, essentially the lifeblood of any business, sales forecasting is becoming crucial in every shopping center. Predicting what will work best, creating business plans for markets that will work, and increasing market knowledge are all always beneficial. Before establishing budgets and marketing strategies for the upcoming year, assumptions are frequently made regarding customer acquisition, funding availability, and strength. Regular sales forecasting research can aid in an in-depth analysis of pre-existing conditions and conditions.

Stated differently, sales forecasts are anticipated using the previous services that are in place. No matter the situation, particularly the external environment, developing and enhancing market opportunities requires in-depth knowledge of the past because it enables planning for the future demands the company. To forecast long-term sales demand, a great deal of research is being conducted in the retailer's industry. The conventional method, also known as the mathematical method, is a significant and efficient way to forecast sales; however, the sales prediction time of these methods is longer. Furthermore, these approaches were unable to handle indirect data; as a result, machine learning techniques were employed in traditional methods to address these issues. Large data sets and indirect data are both easily handled by ML techniques.

1.2. PROBLEM STATEMENT

Many malls and big-box stores are making every effort to maintain their competitive edge in the face of growing competition.One needs a model they can rely on to determine the different factors influencing BigMart's sales as well as the strategies to implement in order to increase profits.Thus, it is possible to create a predictive model that could aid in obtaining important information and boosting revenue.

1.3. OBJECTIVES

Objectives of these project are:

* Predicting future sales from a given dataset.
* To understand the key features that are responsible for the sale of a particular product.
* Find the best algorithm that will predict sales with the greatest accuracy.

1.4. METHOLODY

Represents the steps of building a model. Following are the steps which one needs to follow while creating a model.

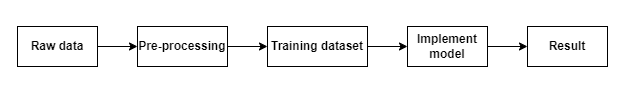


Figure 1.1: Process of building a model

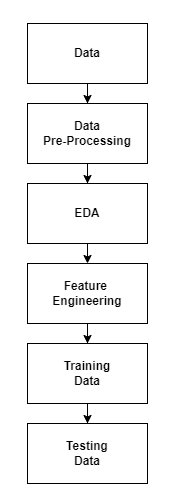


Figure 2.2:Working procedure of proposed model

1.Data collection

The step of every project is to collect the data.

https://www.kaggle.com/datasets/shivan118/big-mart-sales-prediction-datasets/data

2.Data preprocessing

In this step we basically clean our dataset for example check for any missing value in the dataset , if present then handle the missing values. In our dataset attributes like Item Weight and Outlet Size had the missing value.

3.EDA

This part is considered as one of the most important parts when it comes to data analysis.To gain important insights of our data one must need to do exploratory data analysis.Here in our project we used two libraries i.e. klib and dtale library.

4.Tested various algorithms

Then various algorithms like simple LR, xgboost algorithm were applied in order to find out which algorithm can be used to predict the sales.

5.Building the model:

After completing all the previous phases which are mentioned above, now our dataset is ready for further phases that is to build the model. Once we built the model now it is ready to be used as a predictive model to forecast sales of Big Mart.

### ALGORITHMS EMPLOYED

* + 1. **LINEAR REGRESSION (LR)**

As we know Regression can be termed as a parametric technique which means we can predict a continuous or dependent variable on the basis of a provided datasets of independent variables.

The Equation of simple LR is:

Y = βo + β1X + ∈ (1)

where,

Y : It is basically the variable which we used as a predicted value. X : It is a variable(s) which is used for making a prediction.

βo : It is said to be a prediction value when X=0.

β1 : when there is a change in X value by 1 unit then Y value is also changed. It can also be said as slope term ∈

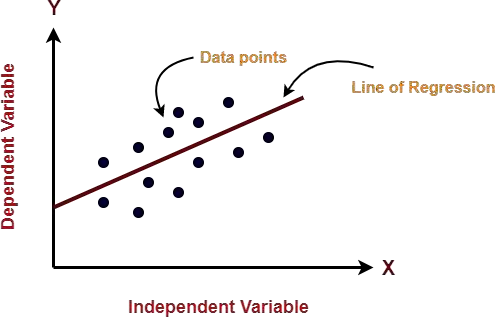


Fig 3.1 Given figure represent line of regression

### RANDOM FOREST REGRESSION

Random Forest is a tree-based bootstrapping algorithm based on that tree that includes a certain number of decision trees to build a powerful predictive model. Individual learners, a set of random lines and a randomly selected few variables often create a tree of choice. The final prediction may be the function of all predictions made by each learner. In the event of a regression. The final prediction may be the meaning of all the predictions.

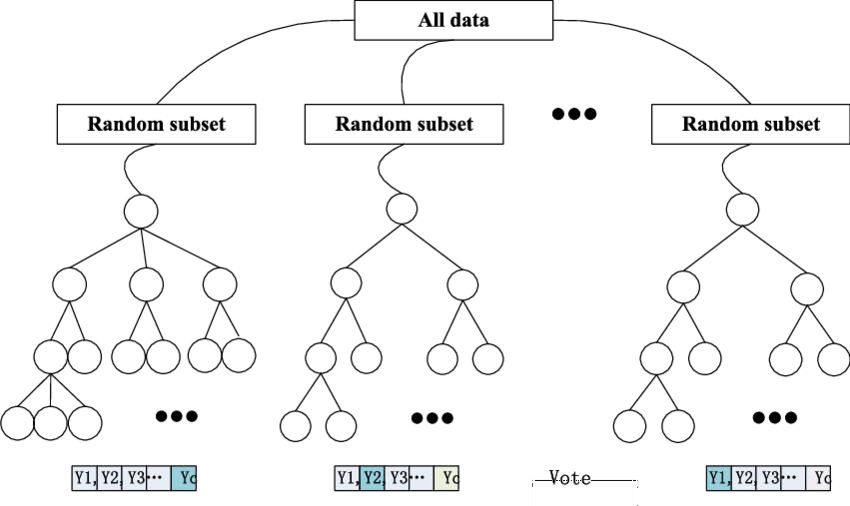


Fig 3.2 : Flowchart of Random Forest Regression

### HYPER PARAMETER TUNING

In ML, optimization of the hyperparameter or problem solving by selecting the correct set of parameters for the learning algorithm. To control the learning process a hyperparameter parameter value is used.In contrast, the values of some parameters are calculated.

The same type of ML model may require different types of weights, learning scales or constraints in order to make different data and information patterns more general. The steps are also called hyperparameters and must be used for the model to solve the ML problem.

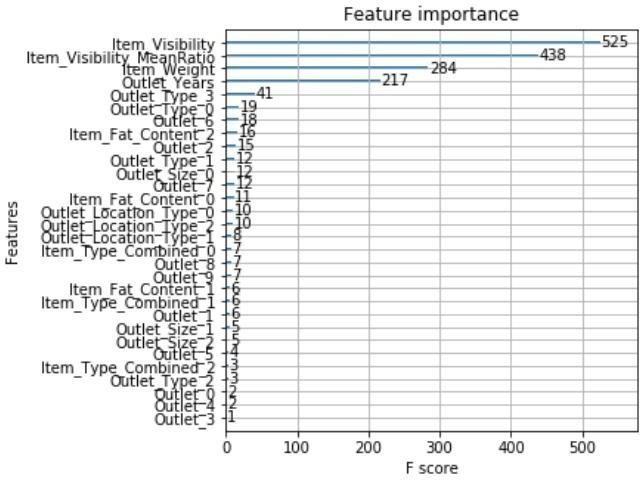


Fig 3.3: Relationship between Feature Importance and their F score in Hyper parameter tuning

### XGBOOST REGRESSION

XGBoost stands for eXtreme Gradient Boosting. The implementation of an algorithm designed for the efficient operation of computer time and memory resources. Boosting is a sequential process based on the principle of the ensemble. This includes a collection of lower learners as well improves the accuracy of forecasts.No model prices n heavy for any minute t, based on the results of the previous t-speed. Well-calculated results are given less weight, and the wrong ones are weighed down. With this algorithm system

The XGBoost model uses stepwise, ridge regression internally, automatically selecting features as well as deleting multicollinearity.

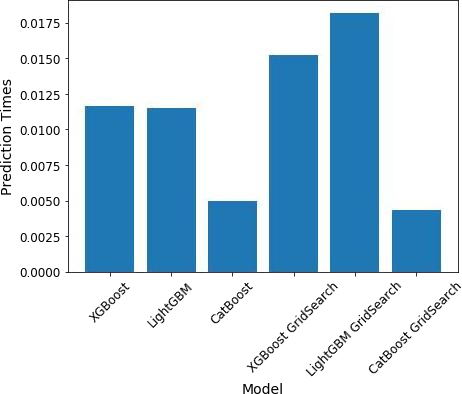


Fig 3.4 : Represents the types of XGBoost regression

* + 1. **DATA AND ITS PREPROCESSING**

In our work, we have used the 2013 Big Mart sales data as a database. Where the data set contains 12 features such as Item Fat, Item Type, MRP Item, Output Type, Object Appearance, Object Weight, Outlet Indicator, Outlet Size, Outlet Year of Establishment, Type of Exit, Exit Identity, and Sales. In these different aspects of responding to the Item Outlet Sales features as well, the other features are also used as the predictive variables. Our dataset has in total 8523 products in various regions and cities. The data set is also based on product level and store- level considerations . Where store level includes features such as city, population density,

store capacity, location, etc. and product-level speculation involves factors such as product, ad, etc. After all considerations, a data set is finally created, then the data set is split into two parts that are tested and trained in a ratio of 80:20.



Fig 3.5: Depicting the features of the dataset