HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY


**MACHINE LEARNING (IT3190E)**

---

# PROJECT: BIGMART SALES PREDICTION

Instructors: Assoc Prof. Khoat Tran
TA Quang Hieu Pham

Students: Nguyen Quang Minh 20204884
Le Van Hoang 20204922
Luong Hoang Anh 20204868
Duong Tung Giang 20204908

Ha Noi, July 2022

# Table of contents

***Abstract***

*In this paper, we will use the prediction of bigmart sales using exploratory machine learning techniques implemented. In a nutshell, sales forecasting is vital to all advertising, marketing, retail, wholesale, and manufacturing operations. This is done by different companies and businesses. This recommendation system will allow businesses to strategize more effectively, achieve higher sales revenue and generate better growth in their own future. We use machine learning methods that aim to give results as close to reality as possible. In the system as a whole, efficiency selection as well as data transformation and exploration will play an important role and give a high level of efficiency to the output in terms of accuracy.*

---

## 1. Introduction

Nowadays, shopping centers and bigmarts are always interested in the sales data of each item to forecast the demand and how much influence that item will have on users in the future and adjust inventory management. In the data warehouse, these data stores contain a substantial amount of consumer information and specific item details. By mining the data store from the datastore, we can detect more anomalies and common patterns can be discovered.

In our research paper, we use the available data set taken from kaggle to analyze and make specific predictions. In this dataset, we are basically provided with specific information about item types such as identifier, weight, fat content,...

With the above problem and data set, we will give practical solutions to solve optimally and effectively through data preparation and application of models in machine learning such as: Linear Regression, Decision Tree and Random Forest. With these models, it will be easy and fast for us to analyze data, visualize graphs, and build machine learning models.

In this report, we will outline the problem solving process. To begin with, we dig through and identify the problem through the available documentation. Next, we use the available data set and focus on mining and analyzing in detail. Finally, by applying machine learning models, we will come up with a solution whose main goal is to evaluate the properties of our data set and make sales predictions at particular stores.

## 2. Problem formulation

To get started in solving any machine learning problem, we always need to define the exact problem we need to solve in order to figure out the optimal methods and strategies and find the best solutions. suitable model.

The problem we studied was a supervised learning problem, so we needed to identify the information of each product to predict the sales of each product type at a particular store.

We define the formula with the trio of Task, Experience and Performance

- Task(T): Predict sales of each type of product at a particular store.
- Experience(E): List of sales data for product categories with certain attributes and defined stores.
- Performance(P): Compare the effectiveness of the models by evaluating their errors.

With the aim of building a predictive model and predicting the sales of each product at a particular store, we have explored methods of data processing and use of valid application models. At the same time, we also know the data processing process of the algorithms used.

# 3. Data preprocessing

## 3.1. Datasets

The Big Mart sales prediction dataset that we use is derived from kaggle. This dataset was loaded by data scientists BigMart collected 2013 sales data of 1559 products across 10 stores in different cities. Furthermore, certain attributes of each product and store were identified. The main purpose is to build a predictive model and predict the sales of each product at a particular store.

Using this model, BigMart tried to understand the product and store characteristics that play an important role in increasing sales.

A few caveats we received from the data set supplier on it was that the data could be missing values as some stores might not be reporting all data due to technical issues. Therefore, they will be required to provide suitable solutions.

** Data dictionary:*

We have data sets train(8523), test(5681) and containing 12 attributes, the dataset train contains both input and output variables.

| Variable | Description |
|---|---|
| Item*Identifier* | Unique product ID |
| ItemWeight | Weight of product |
| Item*Fat*Content | Whether the product is low fat or not |
| Item*Visibility* | The % of the total display area of all products in a store allocated to the particular product |
| ItemType | The category to which the product belongs |
| Item*MRP* | Maximum Retail Price (list price) of the product |
| OutletIdentifier | Unique store ID |
| Outlet*Establishment*Year | The year in which the store was established |
| Outlet*Size* | The size of the store in terms of ground area covered |
| OutletLocation*Type* | The type of city in which the store is located |
| OutletType | Whether the outlet is just a grocery store or some sort of supermarket |
| Item*Outlet*Sales | Sales of the product in t particular store. This is the outcome variable to be predicted. |

*Table 1. Attributes Description*

*Train file:* CSV includes the item outlet information with a sales value.

*Test file:* CSV includes item outlet combinations for which sales need to be forecasted.

### 3.2. Data Exploration and Visualization

An important part of machine learning's problem solving is data visualization. We will give specific views through the analysis of each chart.

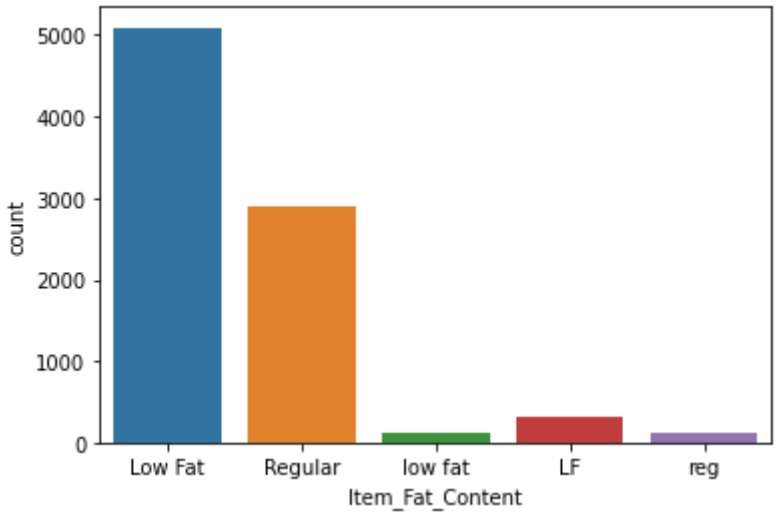


*Figure 1. Item_Fat_Content*

From the chart above, we can directly see that the amount of items that have low fat content (labeled "Low Fat", "LF", "low fat") are much higher than regular fat content (labeled "Regular", "reg"). This shows that customers are always interested in low-fat items, supermarkets and shopping centers are always interested in this and offer products suitable for users. Labeling items with low fat value information is more appealing to customers. From there, business owners and trade centers can make positive predictions to increase product sales.

The next attribute that we analyze is Item_type. According to the description of the bar chart, we provide analysis in 3 groups: The first group is Foods and Drinks (Dairy, Soft Drinks, Meat, Fruits and Vegetables, Baking Goods, Snack Foods, Frozen Foods, Breakfast, Hard). Drinks, Canned, Breads, Starchy Foods, Seafoods), the second group is Houseware (Household, Health and Hygiene) and the third group is Other (others).

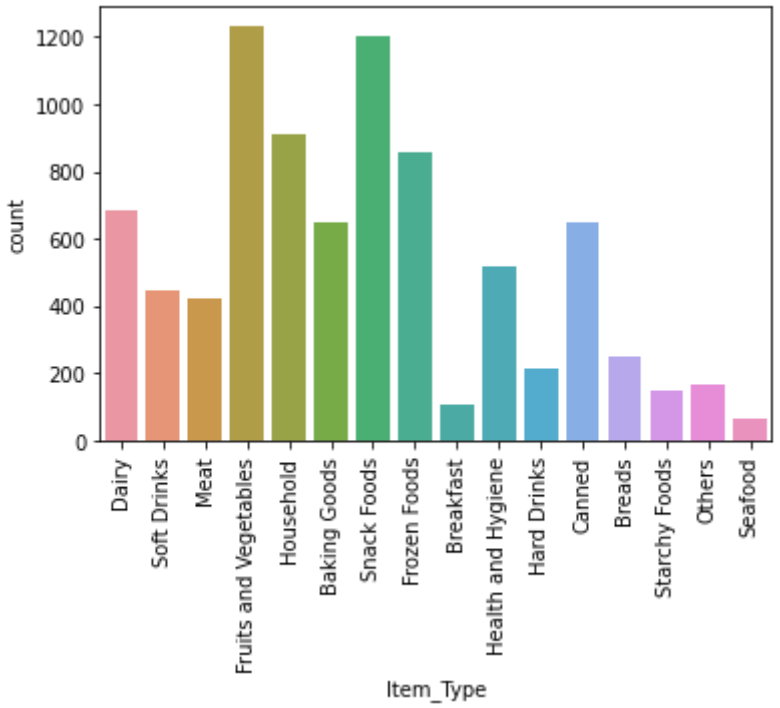


*Figure 2. Item_ Type*

It is easy to see that the Foods and Drinks group contains the most items. Not only that, the number of items in group one is also extremely large and many times larger than the other two groups, especially the two items with the largest quantity, Fruits and Vegetables and Snack Foods, are in this category. Next, the second group (Houseware) contains only two items but the number of these two items is also relatively high. Through the above comments, we can realize the predictions that help businesses and supermarkets to focus on exploiting customers from which types of products.

The third property that we analyze is Outlet_Size. The number of medium-sized stores is 2.5 times larger than that of high stores (approximately 1000 stores). The number of small stores also accounts for a certain number, not inferior to the large stores (medium) (about 2300 stores). From that, it can be guessed that the development level of this city is quite good, the establishment of large stores (medium) at the time of 2013 shows that the consumption level of customers here is quite high.

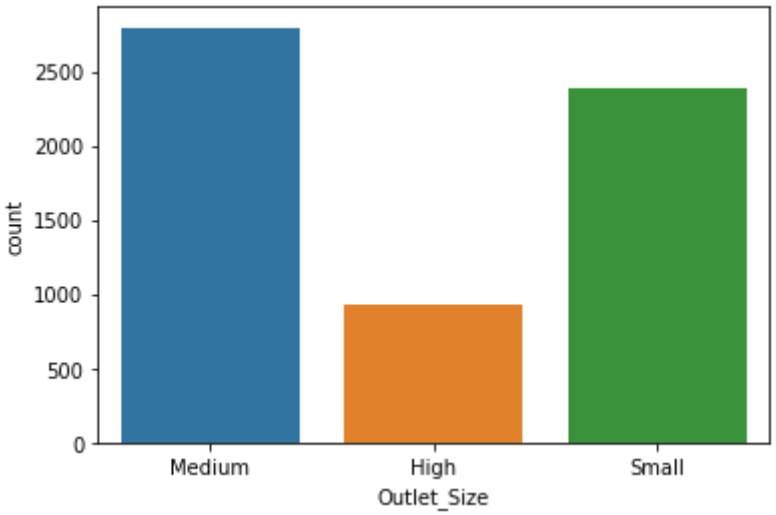


*Figure 3. Outlet_Size*

We can clearly see that Tier 3 type of location is slightly higher than Tier 1 and Tier 2. The Tier 3 location type is close to 3500, followed by Tier 2 near 2800 and Tier1 near 2500.

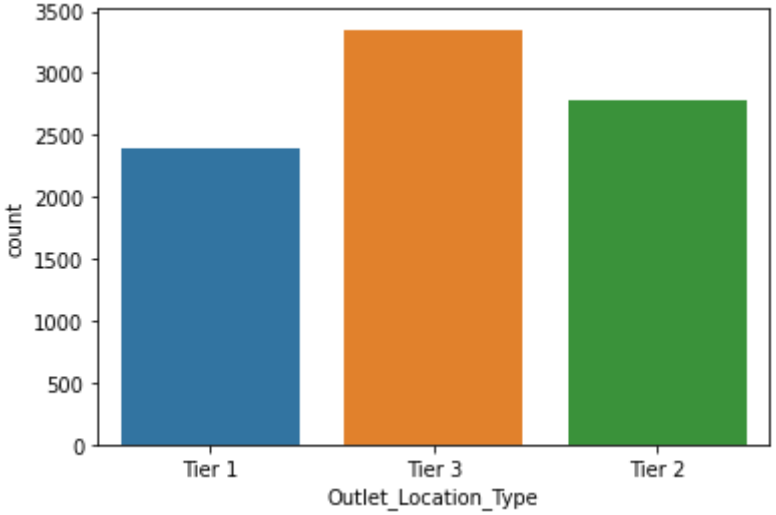


*Figure 4. Oulet_Location_Type*

Supermarket Type 1 is clearly the dominant type of Outlet Type as we can infer from this plot compared with Supermarket Type 2, Supermarket Type 3 and Grocery Store. Through observation, we can see that Supermarket Type 1 is higher than the sum of the remaining three items. Supermarkets Type 2, Supermarkets Type 3 and Grocery Stores have almost the same number (about 1000). The development of Supermarket Type 1 also partly helps to develop trends and increase sales of all kinds of items when consumed here. And that's all for analyzing univariate variables. Next, we'll look at the closer relationship of each predictor to our target variable.

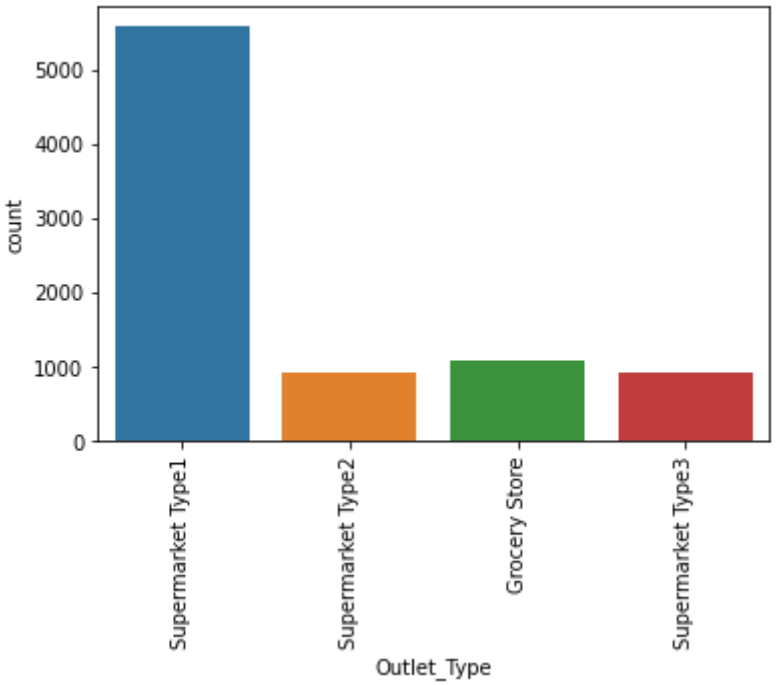


*Figure 5. Outlet_Type*

And that is all for the analysis of univariate variables. Next, we will look closer at the relationship of each predictor to our target variable. First is Item_Weight with Item_Outlet_Size.

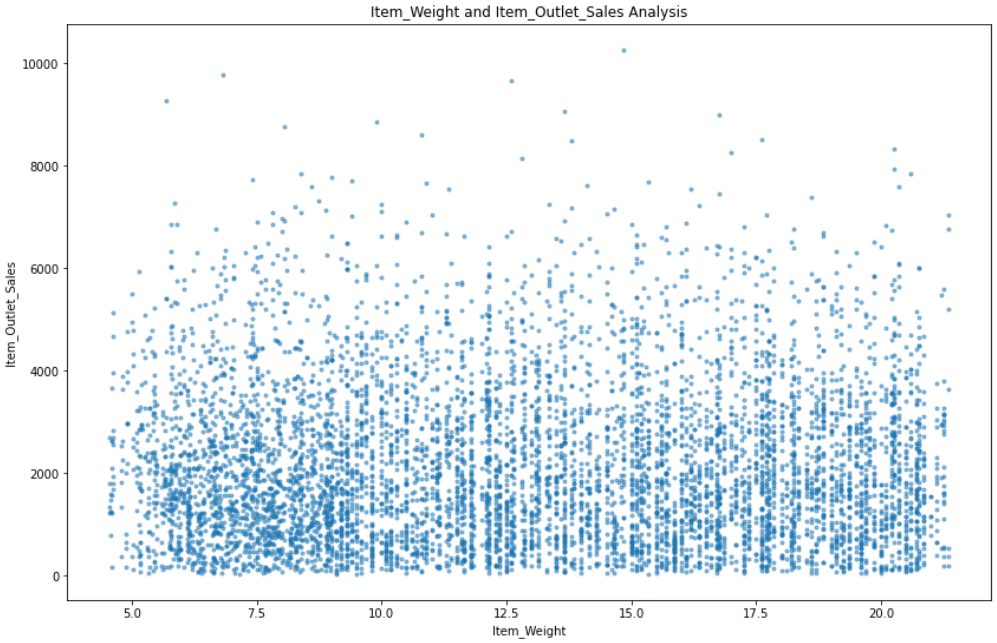


*Figure 6. Item_Weight*

In the source code, we saw that Item_Outlet_Sales and Item_Weight had low correlation. From the plot we can clearly see that this is correct as every value of Item_Weight can have high and low Item_Outlet_Sales. The impact of Item_Weight had on Item_Outlet_Sales is low.

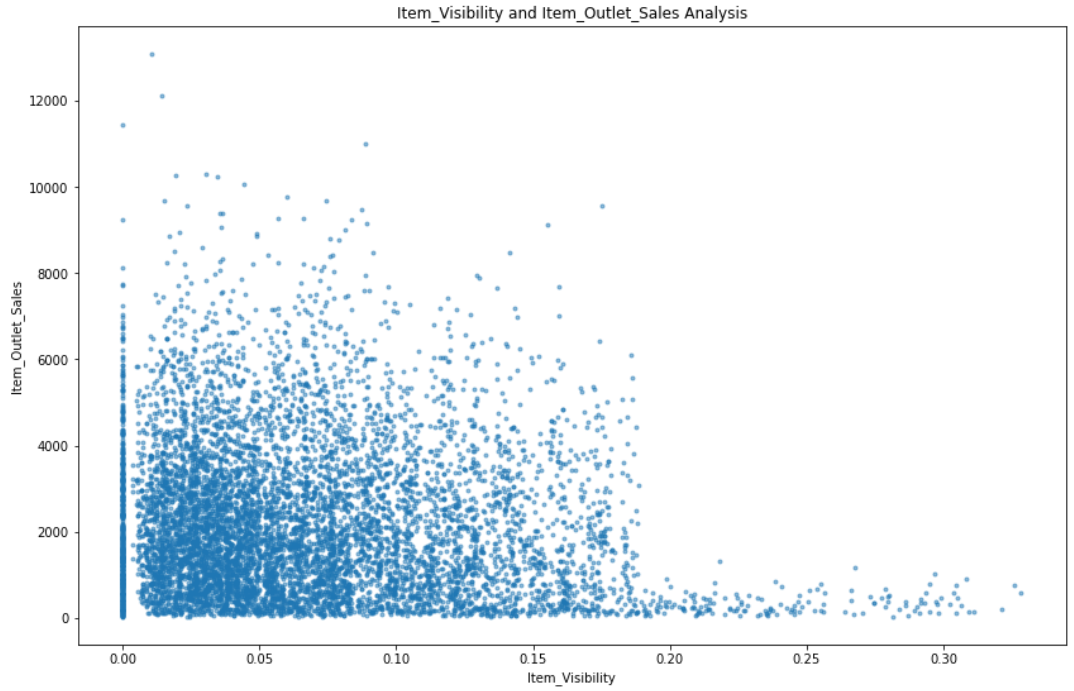


*Figure 7. Item_Visibility*

In this plot, we can see that from 0.20 Item_Visibility onwards the sales seem to be significantly lower. This has been proven to be true since Item_Visibility has negative correlation with Item_Outlet_Sales.

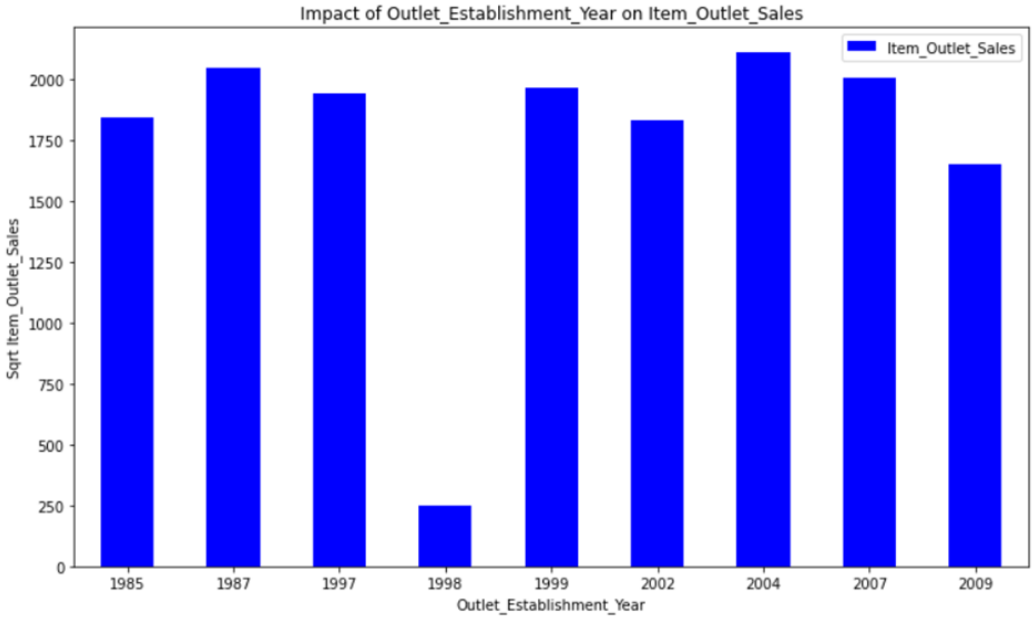


*Figure 8. Outlet_Establishment_Year*

This plot shows that Outlet_Establishment_Year has a low impact on Item_Outlet_Sales as the difference of sales between the years are insignificant.

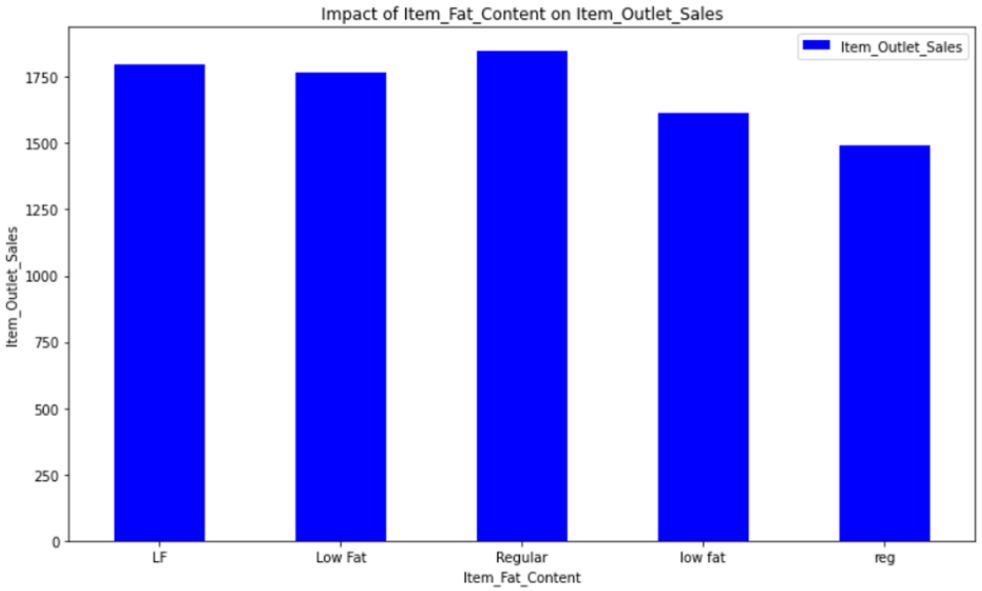


*Figure 9. Item_Fat_Content*

This plot shows that items with low fat content (labeled ‘Low Fat’, ‘LF’, ‘low fat’) seem to have higher sales than regular fat content (labeled ‘reg’, ‘Regular’).

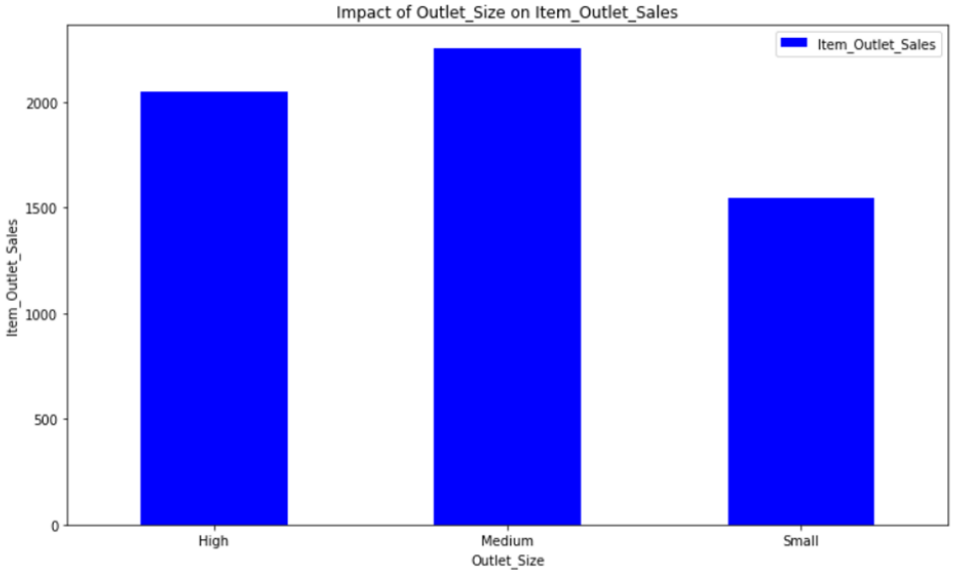


*Figure 10. Outlet_Size*

The impact of Outlet_Size on Item_Outlet_Sales is also an information we need to understand. Medium-sized stores seem to be the dominant type of store compared to high and small sized stores based on Item_Outlet_Sales.

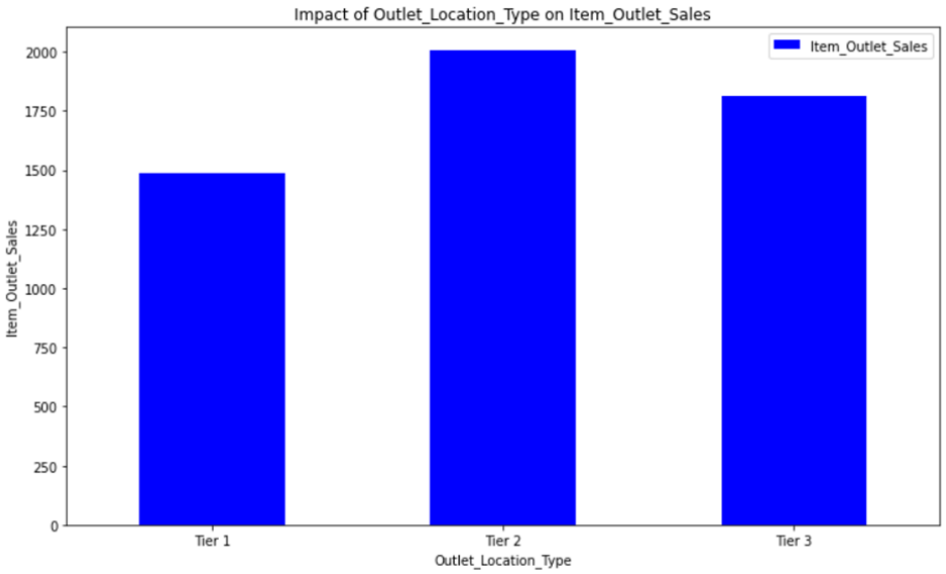


*Figure 11. Outlet_Establishment_Year*

Finally, we will analyze the relationship between Outlet_Location_Type and Item_Outlet_Sales. Tier 1 has the lowest sales among the 3 types of location, followed by tier 2 and lastly, tier 3.

# 4. Model

This section contains the model used for machine learning. We will briefly explain the idea behind it, and evaluate the benefit and drawback of these models.

## 4.1. Linear Regression

A supervised learning method. It estimates the dependent variable Y based on independent variables X by a linear function.

Advantages:

- Linear Regression is a basic method that can be implemented easily.
- It also can be trained easily and efficiently on a system with relatively low computational power when compared with other methods.

Disadvantage:

- The model tends to be underfitting as a linear relationship usually fails to capture real-life data.
- It is also sensitive to outliers (values that deviate when compared to the average mean), ultimately leading to low accuracy.

## 4.2. Decision Tree

Decision Tree are a non-parametric supervised learning method used for classification and regression problems. 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 (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Advantages:

- It can work well without the need of extensive data preparation.
- It is also easy to understand, visualize and have high interpretability.

Disadvantages:

- Decision Tree usually leads to overfitting.
- High training time and cost, and complexity when used with a large dataset.

### 4.3. Random Forest

Random forests or random decision forests is a supervised learning method for classification and regression problems that operate by constructing a multitude of decision trees at training time. It takes the average result of multiple decision trees as the final result.

Advantages:

- Random Forest method usually outperforms Decision Tree.
- It can avoid the overfitting problem of the decision tree by taking the average result.

Disadvantages:

- While random forests often achieve higher accuracy than a single decision tree, they sacrifice the intrinsic interpretability present in decision trees.
- Even higher resource consumption than Decision Tree as it requires several trees to be constructed.

## 5. Experiments

In this section, we will discuss the evaluation metrics used in our problem and why we chose it. We will also present the results obtained from the models used and explain which model is the best.

### 5.1. Evaluation metrics selection

Evaluation matrix is an extremely important stat in evaluating performance or comparing the difference between two models. An important aspect of evaluation metrics is their capability to discriminate among model results.

In some different evaluation metrics such as RMSE, RMLS, R-squared, Adjusted R-Squared, .... we decided to choose 2 stats that are RMSE and R square to evaluate the models in our problem. The formula of RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Where:

- N is the number of data points;
- y(i) is the i-th measurement;
- y^(i) is its corresponding prediction.

The formula of R-squared is:

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

Where:

- MSE(model) Mean Squared Error of the predictions against the actual values;
- MSE(baseline): Mean Squared Error of mean prediction against the actual values.

The reason why we choose RMSE and R-squared to evaluate the models is that both RMSE and $R^2$ quantify how well a regression model fits a dataset. The RMSE tells us how well a regression model can predict the value of the response variable in absolute terms while $R^2$ tells us how well a model can predict the value of the response variable in percentage terms. It's useful to calculate both the RMSE and $R^2$ for a given model because each metric gives us useful information.

## 5.2. Result

After applying the methods and getting the results, here is the table of results obtained with the best parameters we have tried:

| Model | RMSE | R-squared | Parameters |
|---|---|---|---|
| **Linear regression** | 1197 | 0,507893 | |
| **Decision tree** | 1095 | 0,588298 | Max_depth = 15<br>Min_samples_leaf = 300 |
| **Random forest** | 1062 | 0,612598 | Max_depth = 6<br>Min_samples_leaf = 50<br>N_jobs = 4 |

As can be seen from the table above, random forest is the model with the most stable performance when it has the smallest RMSE and the highest R-squared. followed by a decision tree when it has RMSE and R-squared stat in the middle. Finally, linear regression is found to be the least efficient method in our problem while it has the highest RMSE and lowest R-squared.

We can also understand the reason somewhat from the above results. Random forest is a good method. They are based on trees, so scaling of the variables doesn't matter. Any monotonic transformation of a single variable is implicitly captured by a tree. They use the random subspace method and bagging to prevent overfitting. If they are done well, you can have a random forest that deals with missing data easily. Automated feature selection is built in. Linear regression is still an unstable method because of its simplicity.

We also see two evaluation metrics that do not conflict with each other on the results of the evaluation of the models. that shows the stability and efficiency of both stat, especially R-squared.

## 6 Conclusion

In this report, we introduced and solved our prediction problems using different methods in data analysis and machine learning models. Overall, the best model that we conclude is Random Forest, which is the model with the most stable performance through RMSE and R-square evaluation. It has the smallest RMSE and the highest R-square. Along with that, we also found that Linear Regression was the least efficient of the three selected models.

Our biggest obstacle when solving this problem is the selection of the problem model as well as the selection of a strategic process that is reasonable to the problem. In addition, it is relatively difficult to apply a machine learning model to divide the workload fairly for each team member.

In terms of meaning, the study of the topic and the analysis of the given data can bring effective verification, drawing conclusions about the prediction results of each different type of algorithm, from which it is possible to choose the best algorithm for each type of problem. Each research based on existing knowledge is a valuable practical process, contributing to the accumulation of human knowledge, which can then be applied to the real life of each person and bring about positive effects in life.

In order to be able to give each type of model and practical results as above, we had to carefully select the most optimal algorithm types as well as the algorithm with the best value of error. Therefore, if we have more time to research, we will expand the scope of the algorithm and study some new models such as Lasso Regression, Ridge Regression,... along with processing techniques. different from the given data set.

---

## REFERENCES

https://www.kaggle.com/datasets/shivan118/big-mart-sales-prediction-datasets?resource=download

https://en.wikipedia.org/wiki/Linear_regression

https://en.wikipedia.org/wiki/Decision_tree_learning

https://en.wikipedia.org/wiki/Random_forest

https://www.statology.org/rmse-vs-r-squared/

https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/