Internship Report

on

Big mart sales analysis

Submitted by
K R Sindhu
20691A3245

In partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE OR DATA SCIENCE

MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE
(UGC – AUTONOMOUS)

(Affiliated to JNTUA, Ananthapuramu)
Accredited by NBA, Approved by AICTE, New Delhi)
AN ISO 9001:2008 Certified Institution
P. B. No: 14, Angallu, Madanapalle – 517325

2022-2023

✆ : 08571 - 280255, 280590
MADANAPALLE INSTITUTE OF TECHNOLOGY & SCIENCE
AN ENGINEERING COLLEGE SPONSORED BY :
RATAKONDA RANGA REDDY EDUCATIONAL ACADEMY
Approved by AICTE, New Delhi and Affiliated to JNTUA, Anantapur
P.B. No. 14, Angallu, MADANAPALLE - 517 325. Chittoor Dist. A.P.
www.mits.edu

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

BONAFIDE CERTIFICATE

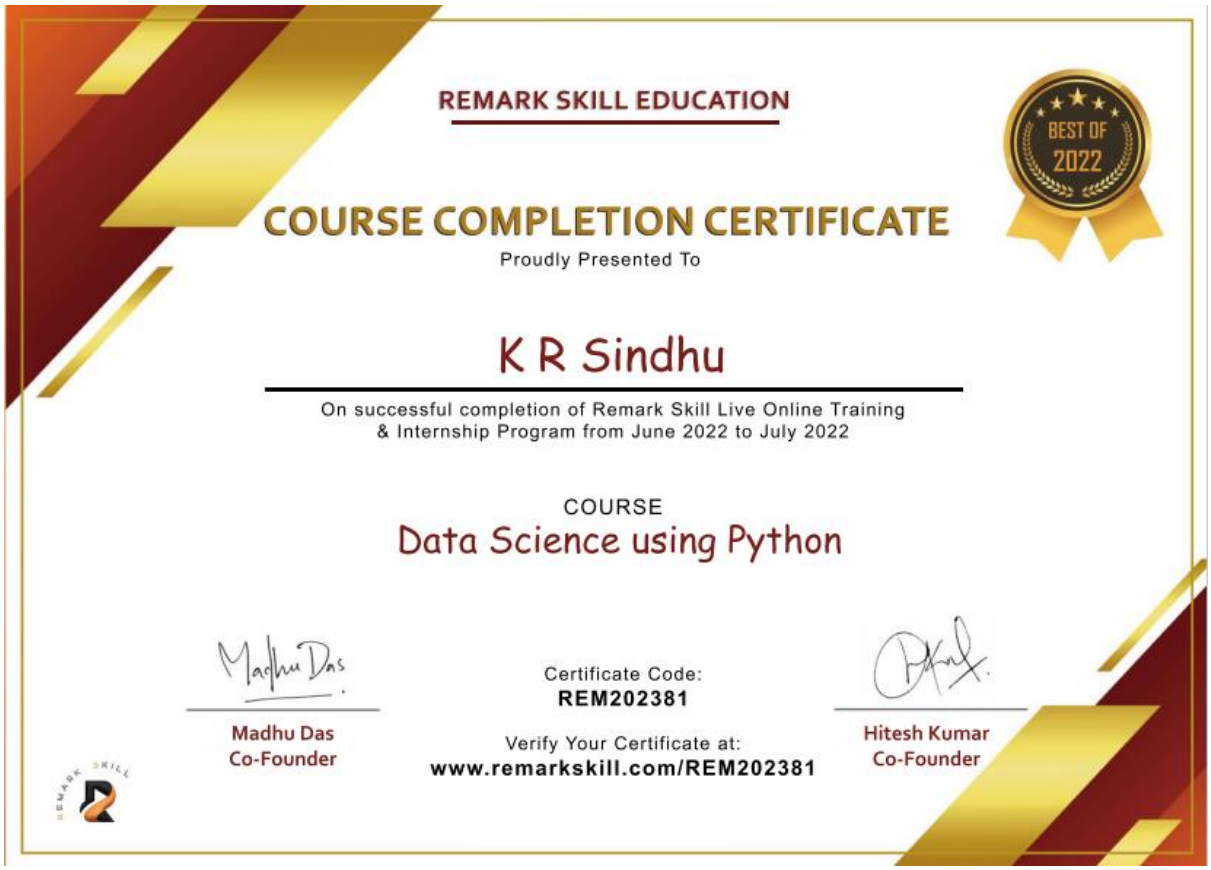
This is to certify that the internship work entitled **Big mart sales analysis** is a bonafide work carried out by

K R Sindhu - 20691A3245

Submitted in partial fulfillment of the requirements for the award of degree **Bachelor of Technology** in the Department of **Data Science**, **Madanapalle Institute of Technology and Science, Madanapalle,** affiliated to **Jawaharlal Nehru Technological University Anantapur, Ananthapuramu** during the academic year 2022-2023

Mrs. Manjula Prabakaran **Assistant Professor** **Department of AI & DS**	**Toralkar Pawan** **Assistant Professor** **Internship Coordinator** **Department of AI & DS**	**Dr. K. Chokkanathan** **Head & Asssociate Professor** **Department of AI & DS**

CERTIFICATE



DECLARATION

I, the undersigned hereby declare that the results embodied in this Internship **"Big mart sales analysis"** is a bonafide record of the work done by me in partial fulfillment of the award of **Bachelor of Technology** in **Data Science** from **Jawaharlal Nehru Technological University Anantapur, Ananthapuramu. The content of this report is** not submitted to any other University/Institute for award of any other degree.

Place:Madanapalle
Date:25-11-2022

K R Sindhu
20691A3245
Department of Artificial Intelligence & Data Science
Madanapalle Institute of Technology & Science
Madanapalle

CONTENTS

ABSTRACT

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, frequent patterns are detected by mining the data 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 project, we propose a predictive model using random forest regression and linear regression technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models. A comparative analysis of the model with others in terms performance metrics is also measured.

ACKNOWLEDGEMENT

I sincerely thank the **Management** of **Madanapalle Institute of Technology and Science** for providing excellent infrastructure and lab facilities that helped me to complete this project.

I sincerely thank **Dr. C. Yuvaraj, M.E., Ph.D**., **Principal** for guiding and providing facilities for the successful completion of our project at **Madanapalle Institute of Technology and Science,** Madanapalle.

I express our deep sense of gratitude to **Dr. K. Chokkanathan, M. Tech., Ph.D.,** Associate Professor & Head, Department of AI & DS for his continuous support in making necessary arrangements for the successful completion of the project.

I express our sincere thanks to the **Internship Coordinator**, **Toralkar Pawan, M.E.,** Assistant Professor, Department of AI & DS for his tremendous support for the successful completion of the internship Project.

.

I express our deep gratitude to **Mrs. Manjula Prabakaran MCA.,MPhil.,M.Tech.,Assistant Professor**, Designation, Department of AI & DS for her guidance and encouragement that helped us to complete this internship project.

I also wish to place on record my gratefulness to other **Faculty members of Department of AI & DS** and also to our friends and our parents for their help and cooperation during our internship.

List of figures

CHAPTER 1
INTRODUCTION

1.1 PYTHON

There are plenty of resources out there that can teach you the basics of machine learning, including books, online courses, and even tutorials on specific algorithms. Once you have a basic understanding of machine learning, you can apply it to real-world problems. In order to make them easy to execute, we need a language that provides vast number of libraries and packages. The presence of this packages reduces the job of a developer by making the execution simple. Among all the 700 languages that are present in the world, python is the best preferred among all.

Introduction to python

Python is a high-level, general-purpose and a very popular programming language. Python programming language (latest Python 3) is being used in web development, Machine Learning applications, along with all cutting edge technology in Software Industry. Python Programming Language is very well suited for Beginners, also for experienced programmers with other programming languages like C++ and Java.

Fig 1.1 symbolic representation of python

Some facts of python programming language:

1. Python is currently the most widely used multi-purpose, high-level programming language.
2. Python allows programming in Object-Oriented and Procedural paradigms.
3. Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.

4. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.
5. The biggest strength of Python is huge collection of standard library.

Python functions:

Python Functions is a block of statements that return the specific task.
The idea is to put some commonly or repeatedly done tasks together and make a function so that instead of writing the same code again and again for different inputs, we can do the function calls to reuse code contained in it over and over again.

Python packages:

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.

Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance & productivity for users.

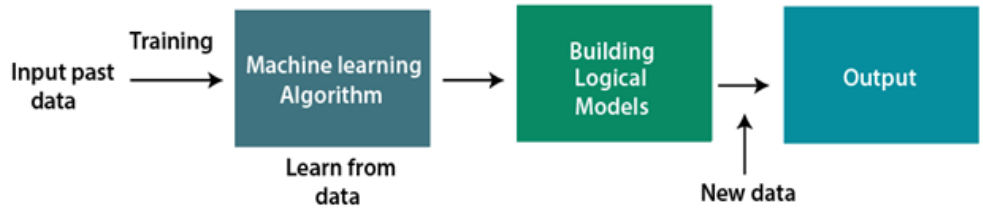
Advantages:

- Fast and efficient for manipulating and analyzing data.
- Data from different file objects can be loaded.
- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Data set merging and joining.
- Flexible reshaping and pivoting of data sets

1.2 Machine learning language

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models

and making predictions using historical data or information**.** Currently, it is being used for various tasks such as image recognition**,** speech recognition**,** email filtering**,** Facebook auto-tagging**,** recommender system and many.



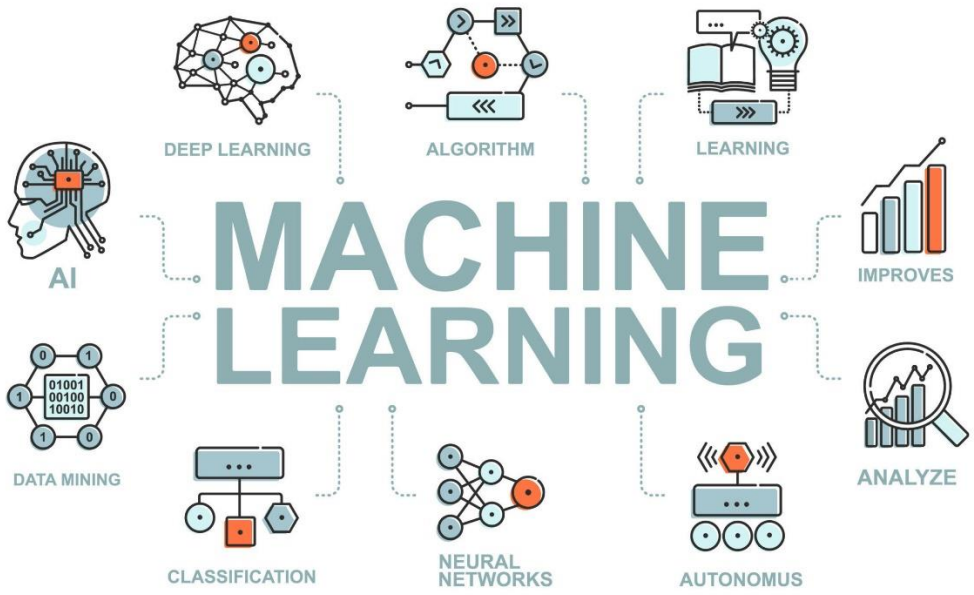


Fig.1.2 Machine learning

Features of Machine Learning:

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Need for Machine Learning

The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to

implement directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

The importance of machine learning can be easily understood by its uses cases, Currently, machine learning is used in self-driving cars**,** cyber fraud detection**,** face recognition**,** and friend suggestion by Facebook, etc. Various top companies such as Netflix and Amazon have build machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**

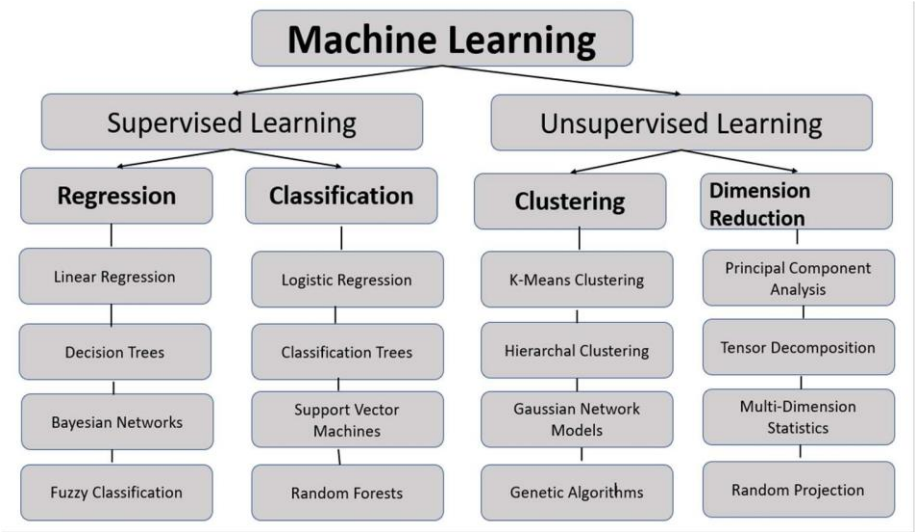


Fig 1.3 classification of machine learning

1.Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output. The supervised learning is based on supervision, and it is the same as when a student learns

things in the supervision of the teacher. The example of supervised learning is spam filtering. Supervised learning can be grouped further in two categories of algorithms:

- **Classification**
- **Regression**

Classification is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some parameters given in input and then the labels are predicted for the data.

Regression:

A regression problem is when the output variable is a real or continuous value, such as "salary" or "weight". Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points.

Some famous Examples of regression problems are:

- Predicting the house price based on the size of the house, availability of schools in the area, and other essential factors.
- Predicting the sales revenue of a company based on data such as the previous sales of the company.
- Predicting the temperature of any day based on data such as wind speed, humidity, atmospheric pressure.

2.Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision.The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns. In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classifieds into two categories of algorithms:

- **Clustering**
- **Association**

3.Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance. The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

Applications of Machine learning

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications of Machine Learning:

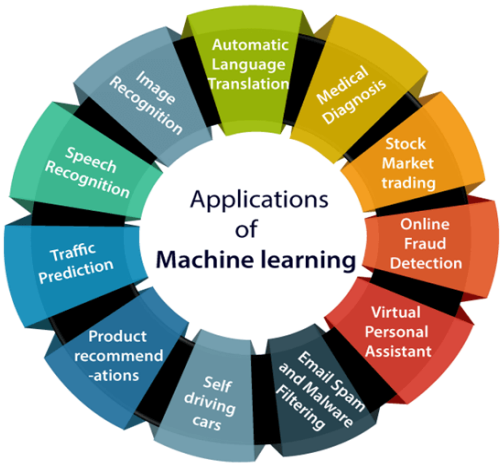


Fig. 1.4 Applications of Machine Learning

1.3 Exploratory data analysis

- **Exploratory Data Analysis (EDA)** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

- It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today**.**
- By the process of univariant analysis we can predict what type of item consumes prefer more, how much fat content is preferable, which items are having more sales.
- Bivariate analysis explores how the dependent ("outcome") variable depends or is explained by the independent variable.

1.4 Plotting techniques

Box Plots

Box and Whisker plots, also called boxplots in short, is another useful technique to review the distribution of each attribute's distribution. The following are the characteristics of this technique –

1.It is univariate in nature and summarizes the distribution of each attribute.

2.It draws a line for the middle value i.e. for median.

3.It draws a box around the 25% and 75%.

4.It also draws whiskers which will give us an idea about the spread of the data.

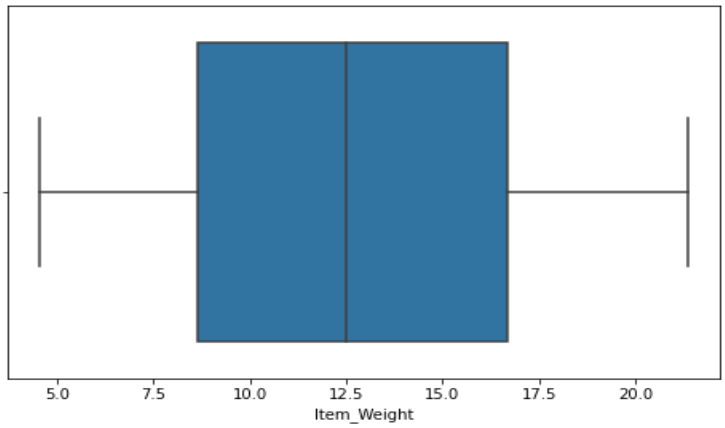


Fig 1.5 Box plot

Scatter Plot

Scatter plots shows how much one variable is affected by another or the relationship between them with the help of dots in two dimensions. Scatter plots are very much like line graphs in the concept that they use horizontal and vertical axes to plot data points.

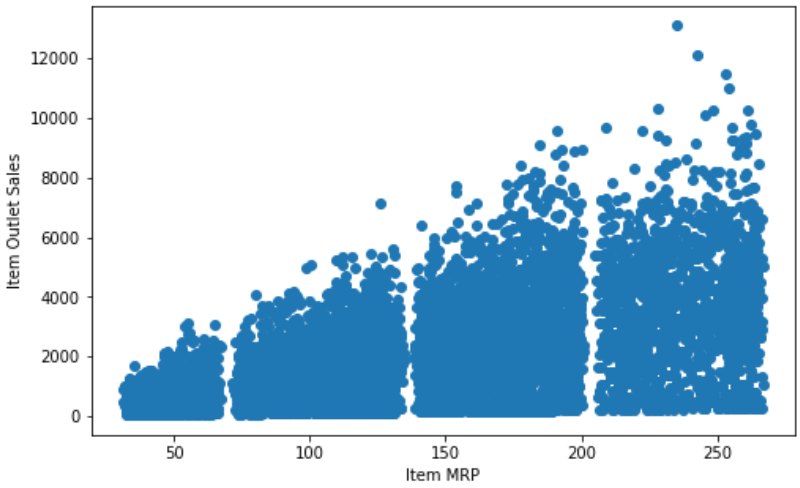


Fig 1.6 Scatter plot

Bar plot

A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

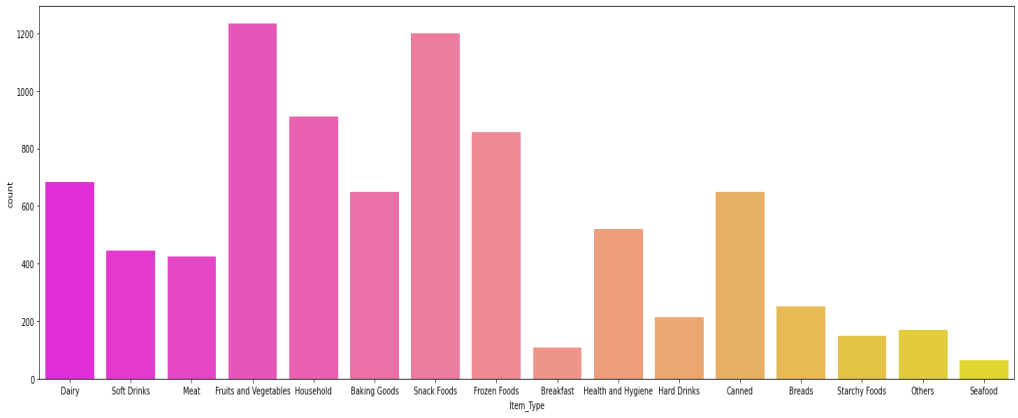


Fig 1.7 Bar plot

CHAPTER - 2

TOOLS AND TECHNIQUES

2.1 HARDWARE REQUIRED

- Device name : HP Laptop 15s-fq4xxx
- Processor : 11th Gen Intel(R) Core(TM) i5-1155G7 @ 2.50GHz 2.50 GHz
- Installed RAM : 8.00 GB (7.65 GB usable)
- System type : 64-bit operating system, x64-based processor

2.2 SOFTWARE REQUIRED

- Google Collab is a free Jupiter Notebook environment that runs entirely in Google Cloud. Moreover, it has many libraries already installed to manipulate data and train Machine Learning models.
- Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education
- More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs.

As a programmer, you can perform the following using Google Collab.

- Document your code that supports mathematical equations
- Import external datasets e.g. from Kaggle
- Free Cloud service with free GPU
- Easy to perform graphical representations such as graphs, pie charts etc.,

Fig 2.1 Symbolic representation of collab

CHAPTER 3

PROJECT ANALYSIS

3.1 PROJECT OVERVIEW

This report is a brief description of our 5 weeks internship on Data Science. The internship was carried out from JUNE 2022 to JULY 2022. The project was carried out using python. And my project was on "BIG MART SALES ". So we use Google Collabs or Jupyter Notebook as the software to import libraries and excecute the program. First insert the data set which we chosen and then start the execution. This model will be useful to the big mart owners to predict their future sales and to gain more profit.It is Just a small Project to Learn the basic concepts of Machine Learning. Though Machine Learning is a sub field of Artificial Intelligence , but it has a vast real life applications .Its an small Field but enough to Build our Future.

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.

Problem statement

In this project, we are addressing the problem of big mart sales prediction or forecasting of an item on customer's future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc are used for prediction or forecasting of sales volume. 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 to improve the knowledge of marketplace.

A standard sales prediction can help in deeply analyzing the situations or the conditions previously occurred and then, the inference can be applied about customer acquisition, and strengths before setting a budget and marketing plans for the upcoming year. In other words, sales prediction is based on the available resources from the past. In depth knowledge of past is required for enhancing and improving the likelihood of marketplace irrespective of any circumstances especially the external circumstance, which allows to prepare the upcoming needs for the business.

Existing system

Das P. Chaudhury done Prediction of retail sales of big mart using feedforward and recurrent Neural Networks (2018) by using neural networks for prediction of sales he could not get more accurate results which is not efficient, So random forest regressor can work efficiently.

Proposed System

On implementation, the prediction results show the correlation among different attributes considered and gives accurate and consistent results.
This model having any number of users can be supported by the system.

3.2 Modules

About the dataset

I collected big mart sales dataset from Kaggle. The data scientists at big mart have collected 2013 sales data for 1559 products across 10 stores in different cities. The dataset contains 12 columns. Also certain attribute of each product and store have been defined.

- Item Identifier : Unique product 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 (list 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 supermarket
- Item Outlet Sales : Sales of the product in the particular store.

Data pre processing

Data pre processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

Why do we need Data Preprocessing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. It involves following steps

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Exploratory data analysis

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Why Is Exploratory Data Analysis Important?

EDA is important as it allows data scientists to analyze the data before coming to any assumptions and ensures that the results produced are valid and applicable to business outcomes and goals. It has the following features:

- Helps identify errors
- Promotes better understanding of patterns within the data
- Helps detect abnormal events.
- Helps understand data set variables and the relationship among them.

Machine learning algorithm

To predict the big mart sales supervised machine learning algorithms are used. Random forest regression and linear regression are mainly used algorithms to predict the sales.

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 on multiple decision trees. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **Aggregation**.

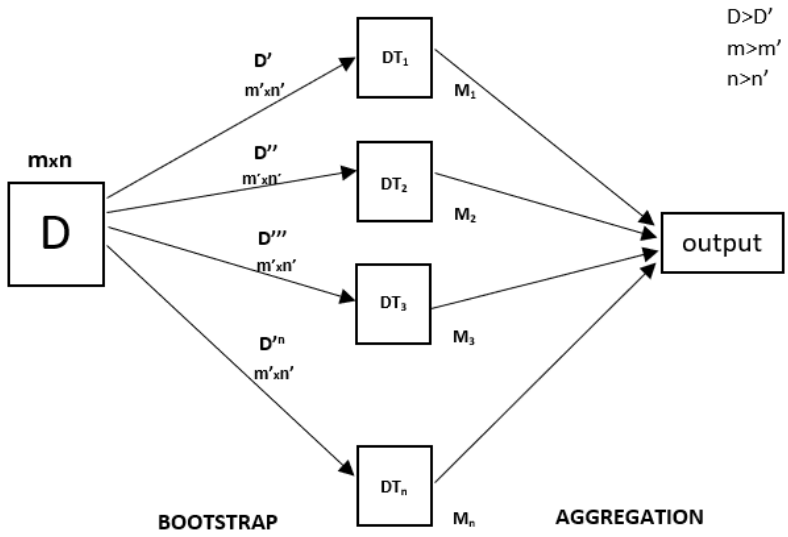


Fig 3.1 Random forest regression

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

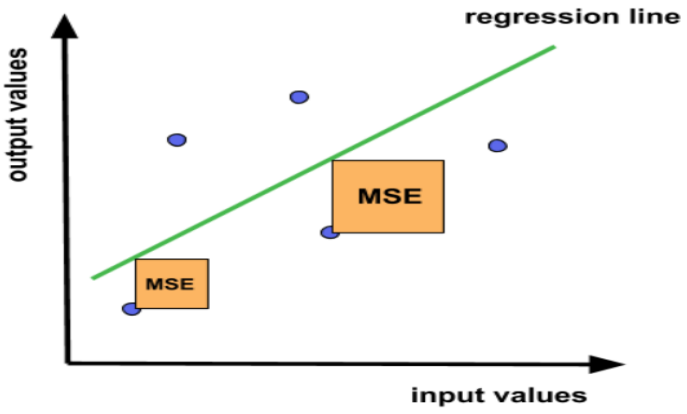
multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models.

We need to approach the Random Forest regression technique in the following way

- Design a specific question or data and get the source to determine the required data.
- Make sure the data is in an accessible format else convert it to the required format.
- Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
- Create a machine learning model
- Set the baseline model that you want to achieve
- Train the data machine learning model.
- Provide an insight into the model with test data
- Now compare the performance metrics of both the test data and the predicted data from the model.
- If it doesn't satisfy your expectations, you can try improving your model accordingly or dating your data, or using another data modeling technique.
- At this stage, you interpret the data you have gained and report accordingly.

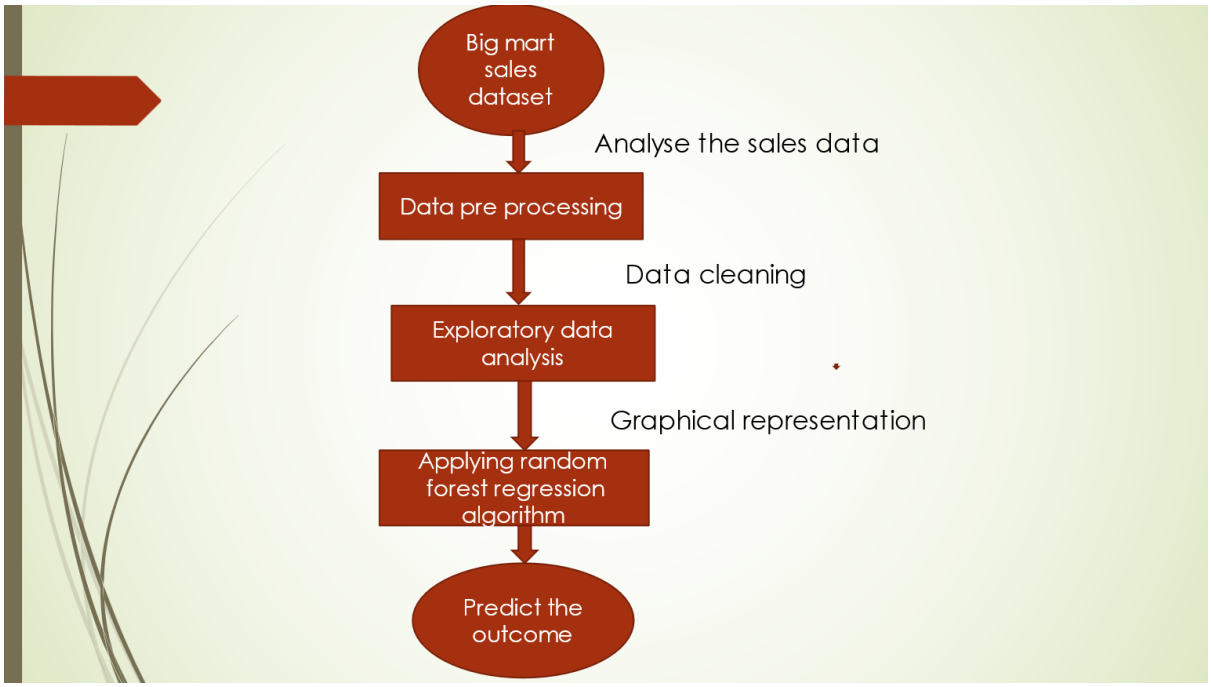
Linear regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task**.** Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.Linear regression performs the task to predict a dependent variable value (y)



based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regressio
Fig 3.2 Linear regression

3.3 Architecture



CHAPTER 4

CONCLUSION

In present era of digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items in all seasons. Analysed the big mart sales data set to predict the future demand and to gain maximise frofit. 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. It can be concluded that more locations should be switched or shifted to Supermarket Type3 to

increase the sales of products at Big Mart. Any one-stop-shopping-center like Big Mart can benefit from this model by being able to predict its items' future sales at different locations. Item MRP clearly maximises the Outlet sales. Outlet Type being Super Market or Grocery Store also depicts the Outlet Sales. And Years Established will also tend to have loyal customers and that's how the sales margin can be acheived.

CHAPTER 5

BIBLIOGRAPHY

- **https://**www.kaggle.com/datasets/brijbhushannanda1979/bigmart-sales-data
- (PDF) A Comparative Study of Big Mart Sales Prediction (researchgate.net)
- Bigmart Dataset Sales Prediction. This post is about my approach on… | by Vishal Borana | Analytics Vidhya | Medium

CHAPTER 6

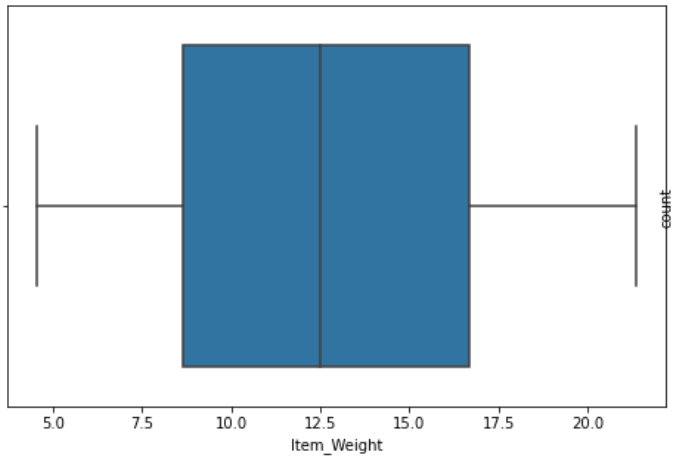
APPENDIX

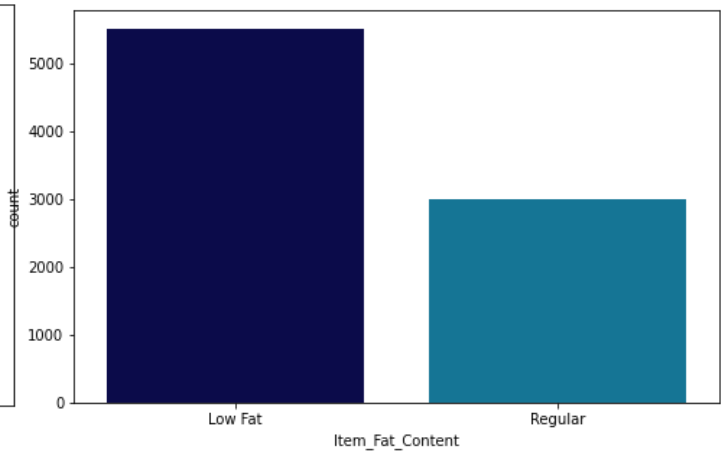
A. Sample code

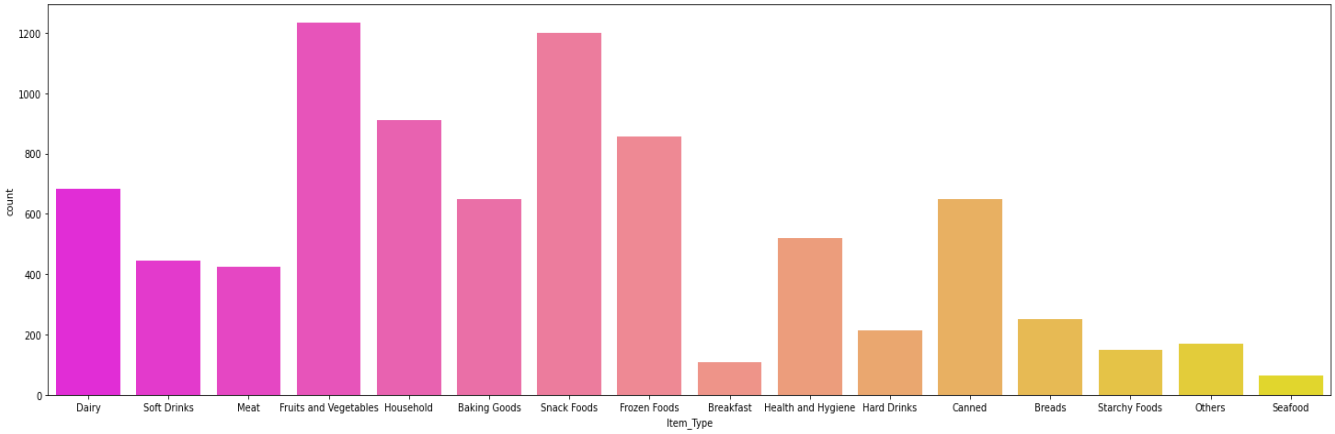
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.model_selection  import cross_val_score
from sklearn import metrics
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
#Load the dataset
train = pd.read_csv("TRAIN.csv")
test = pd.read_csv("TEST.csv")
train.shape,test.shape
#find the missing values
test.apply(lambda x: sum(x.isnull()))
train.apply(lambda x: sum(x.isnull()))
train.info()
categorical = train.select_dtypes(include =[np.object])
print("Categorical Features in Train Set:",categorical.shape[1])
numerical= train.select_dtypes(include =[np.float64,np.int64])
print("Numerical Features in Train Set:",numerical.shape[1])
categorical = test.select_dtypes(include =[np.object])
print("Categorical Features in Test Set:",categorical.shape[1])
numerical= test.select_dtypes(include =[np.float64,np.int64])
print("Numerical Features in Test Set:",numerical.shape[1])
#box plot
plt.figure(figsize=(8,5))
sns.boxplot('Item_Weight',data=train)
plt.figure(figsize=(8,5))
sns.boxplot('Item_Weight',data=test)
#Exploratory data analysis
plt.figure(figsize=(25,7))
sns.countplot('Item_Type',data=train,palette='spring')
plt.figure(figsize=(8,5))
sns.countplot('Outlet_Size',data=train,palette='summer')
plt.figure(figsize=(8,5))
sns.countplot('Years_Established',data=train,palette='mako')
plt.figure(figsize=(10,8))
sns.barplot(y='Item_Type',x='Item_Outlet_Sales',data=train,palette='flag')
plt.figure(figsize=(8,5))
```

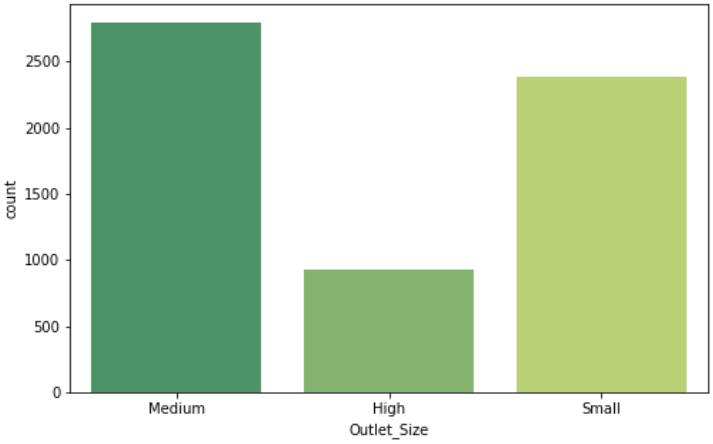
```
plt.scatter(y='Item_Outlet_Sales',x='Item_MRP',data=train)
plt.xlabel('Item MRP')
plt.ylabel('Item Outlet Sales')
plt.show()
plt.figure(figsize=(8,5))
sns.barplot(x='Years_Established',y='Item_Outlet_Sales',data=train,palette='viridis')
#label encoding
le = LabelEncoder()
var_mod = ['Item_Fat_Content','Outlet_Location_Type','Outlet_Size','Outlet_Type','Item_Ty
pe']
for i in var_mod:
    train[i] = le.fit_transform(train[i])
for i in var_mod:
    test[i] = le.fit_transform(test[i])
#test and train the data
X= train[['Item_Weight','Item_Fat_Content','Item_Visibility','Item_Type','Item_MRP','Outlet
_Size','Outlet_Location_Type','Outlet_Type','Years_Established']]
y= train['Item_Outlet_Sales']
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=22)
#Random forest
RF= RandomForestRegressor(n_estimators=200,max_depth=5, min_samples_leaf=100,n_job
s=4,random_state=22)
# Fitting the model on our trained dataset.
RF.fit(X_train,y_train)
# Making Predictions
y_pred = RF.predict(X_test)
coef3 = pd.Series(RF.feature_importances_, features).sort_values(ascending=False)
coef3
plt.figure(figsize=(8,5))
sns.barplot(RF.feature_importances_,features)
#evaluating model
def cross_val(model_name,model,X,y,cv):
    scores = cross_val_score(model, X, y, cv=cv)
    print(f'{model_name} Scores:')
    for i in scores:
        print(round(i,2))
    print(f'Average {model_name} score: {round(scores.mean(),2)}')
```

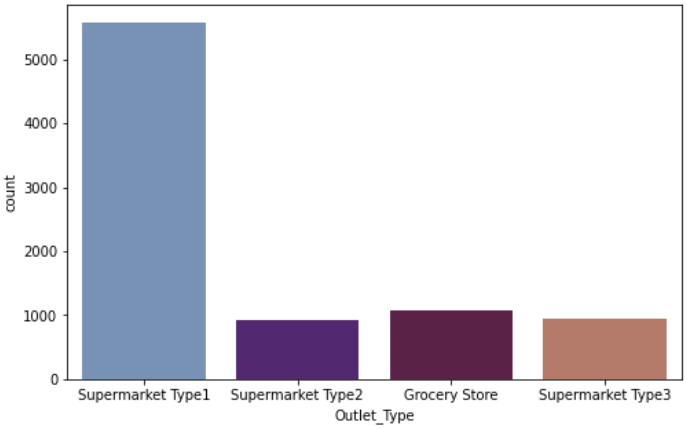
B. Output

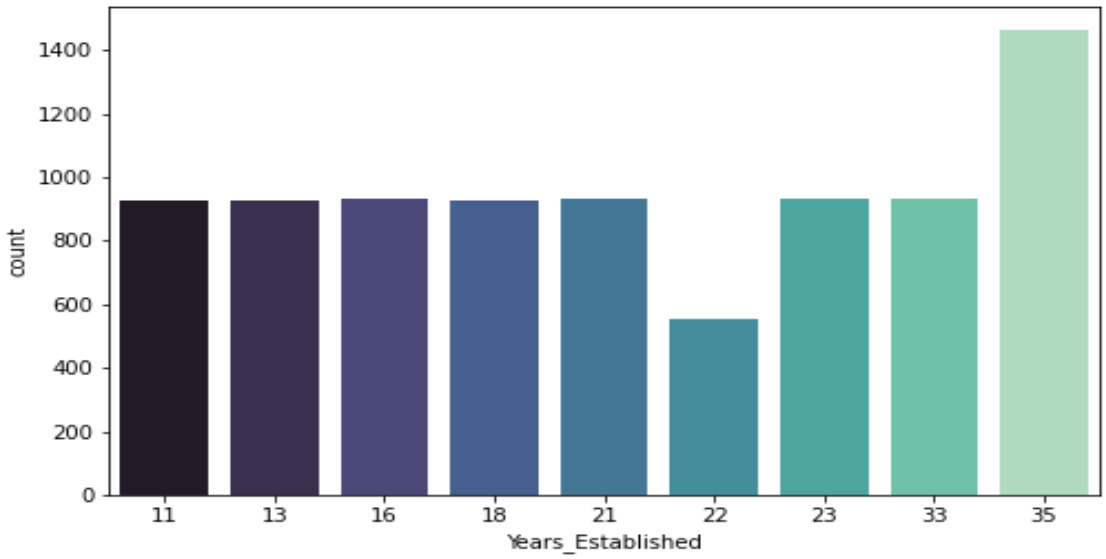

5.0
7.5
10.0
12.5
15.0
17.5
20.0
Item_Weight

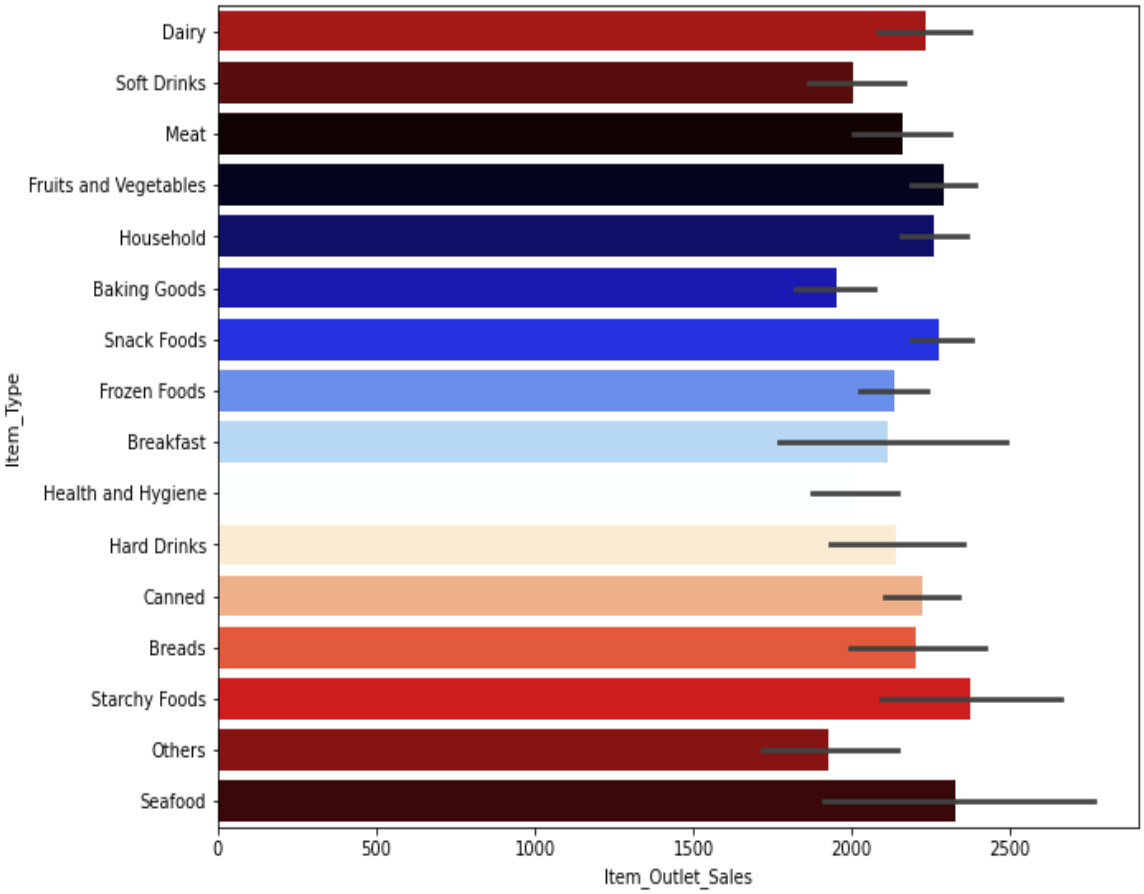

count
5000
4000
3000
2000
1000
0
Low Fat
Regular
Item_Fat_Content

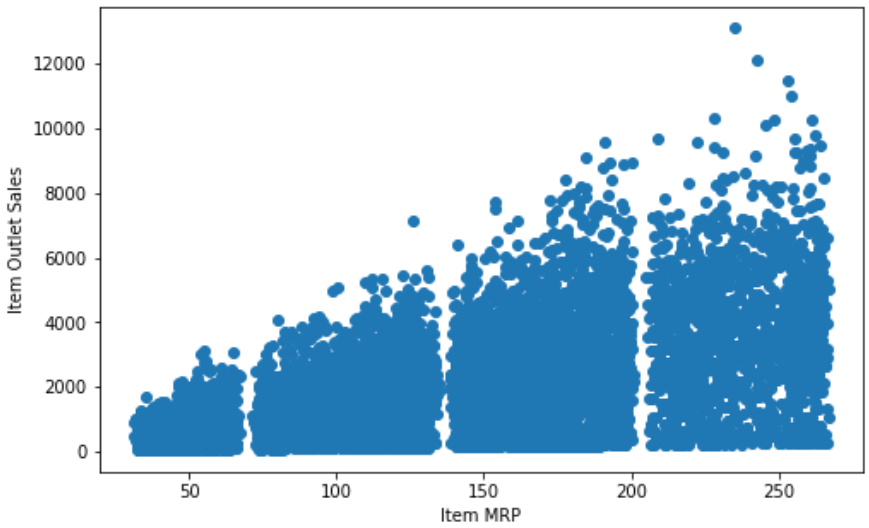

count
1200
1000
800
600
400
200
0
Dairy
Soft Drinks
Meat
Fruits and Vegetables
Household
Baking Goods
Snack Foods
Frozen Foods
Breakfast
Health and Hygiene
Hard Drinks
Canned
Breads
Starchy Foods
Others
Seafood
Item_Type

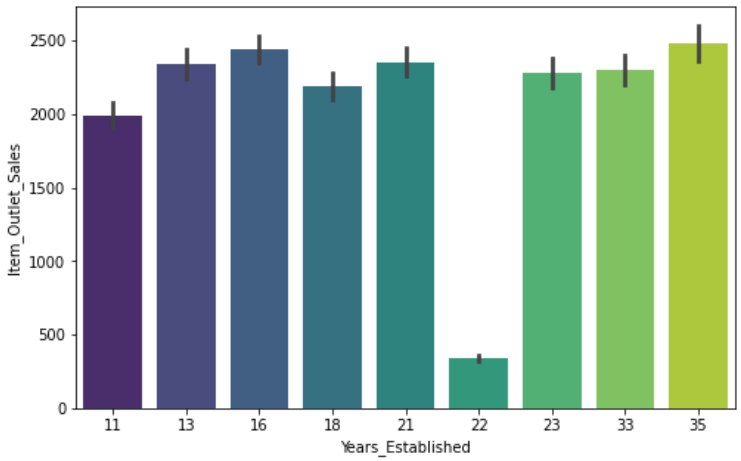

count
2500
2000
1500
1000
500
0
Medium
High
Small
Outlet_Size

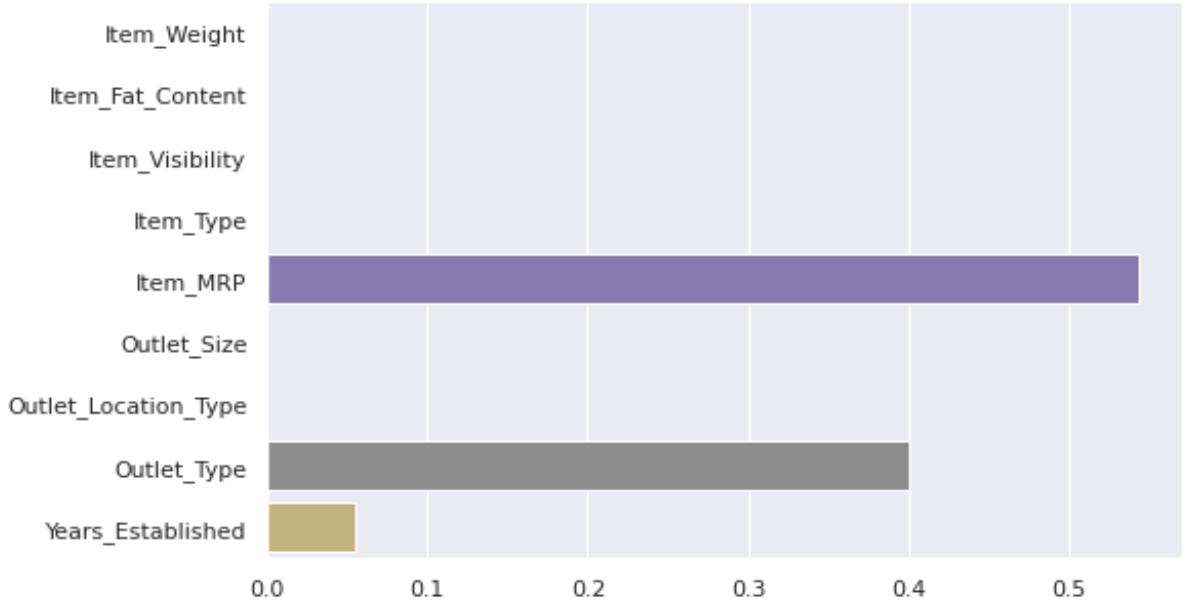

count
5000
4000
3000
2000
1000
0
Supermarket Type1
Supermarket Type2
Grocery Store
Supermarket Type3
Outlet_Type


count
1400
1200
1000
800
600
400
200
0
11
13
16
18
21
22
23
33
35
Years_Established


Item_Type
Dairy
Soft Drinks
Meat
Fruits and Vegetables
Household
Baking Goods
Snack Foods
Frozen Foods
Breakfast
Health and Hygiene
Hard Drinks
Canned
Breads
Starchy Foods
Others
Seafood
0
500
1000
1500
2000
2500
Item_Outlet_Sales


Item Outlet Sales
12000
10000
8000
6000
4000
2000
0
50
100
150
200
250
Item MRP


Item_Outlet_Sales
2500
2000
1500
1000
500
0
11
13
16
18
21
22
23
33
35
Years_Established


Item_Weight
Item_Fat_Content
Item_Visibility
Item_Type
Item_MRP
Outlet_Size
Outlet_Location_Type
Outlet_Type
Years_Established
0.0
0.1
0.2
0.3
0.4
0.5

```
Item_MRP                0.543841
Outlet_Type             0.400777
Years_Established       0.054463
Item_Visibility         0.000263
Outlet_Location_Type    0.000248
Item_Weight             0.000241
Outlet_Size             0.000087
Item_Type               0.000078
Item_Fat_Content        0.000000
dtype: float64
```

```
RandomForestRegressor(max_depth=5, min_samples_leaf=100, n_estimators=200,
                      n_jobs=4, random_state=22) Scores:
0.57
0.54
0.52
0.55
0.57
Average RandomForestRegressor(max_depth=5, min_samples_leaf=100, n_estimators=200,
                      n_jobs=4, random_state=22) score: 0.55
```