A

Project Report

On

Sales Forecasting using ML

B. Tech (Electronics and Telecommunication Engineering)

Submitted By

Aman Jain (70061118047)

Aritro Sengupta (70061118037)

Dhrumil Vadgama (70061118044)

Under the Guidance of

Prof. Sunil Chaudhari

Department of Electronics and Telecommunication Engineering

S.V.K.M's NMIMS, Mukesh Patel School of Technology

Management and Engineering.

Academic Session-2021-2022

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION
Mukesh Patel School of Technology Management & Engineering

CERTIFICATE

This is to certify that Project titled **" Sales Forecasting using ML "** has been successfully completed by

Aman Jain (70061118047)

Aritro Sengupta (70061118037)

Dhrumil Vadgama (70061118044)

Under the guidance of

Prof. Sunil Chaudhari

in partial completion of the requirement for Bachelor degree in Electronics of MPSTME, SVKM's NMIMS University, Mumbai, Shirpur Campus during the academic year 2021-2022.

Date: 26th March, 2022

Place: Shirpur

Prof. Sunil Chaudhari
Project (Mentor)

Prof. Atul Patil
Head,
Department of Electronics & Telecommunication Engineering

Dr Kamal Mehta
Associate Dean
(MPSTME-NMIMS, Shirpur Campus)

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION
Mukesh Patel School of Technology Management & Engineering

CERTIFICATE FOR APPROVAL

The Project titled **Sales Forecasting using ML** being submitted by

Aman Jain (70061118047)

Aritro Sengupta (70061118037)

Dhrumil Vadgama (70061118044)

has been examined by us and is hereby approved for the award of degree "BACHELOR OF TECHNOLOGY in Electronics & Telecommunication Engineering" discipline for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approved the project only for the purpose for which it has been submitted.

Place: Shirpur

Internal Examiner **External Examiner**

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION
Mukesh Patel School of Technology Management & Engineering

DECLARATION

We,

Aman Jain

Aritro Sengupta

Dhrumil Vadgama

the students of **Bachelor of Technology** in **Electronics and Telecommunications discipline, Session: 2021-22, MPSTME, Shirpur Campus,** hereby declare that the work presented in this project entitled **"Sales Forecasting using ML"** is the outcome of our work, is bona fide and correct to the best of our knowledge and this work has been carried out taking care of Engineering Ethics. The work presented doesn't infringe any patented work and hasn't been submitted to any other university or anywhere else for the award of any degree or any professional diploma.

Aman Jain (70061118047)

Aritro Sengupta (70061118037)

Dhrumil Vadgama (70061118044)

Date : 26th March, 2022

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION
Mukesh Patel School of Technology Management & Engineering.

ACKNOWLEDGEMENT

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organization. We would like to extend my sincere thanks to all of them.

Thank my god for providing us with everything that we required in completing this project.

We are highly indebted to the teacher in charge Prof. Pankaj Gulhane for his constant guidance and impeccable supervision as well as for providing necessary information regarding the project and also for his support in completing the project.

We would like to express my gratitude towards my parents for their kind co-operation and encouragement which helped us in the completion of this project.

We would like to express my special gratitude and thanks to industry persons for giving us such attention and time.

Our thanks and appreciations also go to our classmates and batch mates in developing the project and to the people who have willingly helped us out with their abilities.

Aman Jain (70061118047)

Aritro Sengupta (70061118037)

Dhrumil Vadgama (70061118044)

ABSTRACT

Forecasting sales is a common and essential use of machine learning (ML). Sales forecasting is both a science and an art. Decision makers rely on these forecasts to plan for business expansion and to determine how to fuel the company's growth. So, in many ways, sales forecasting affects everyone in the organization.

Sales forecasts are typically based on data collected over the years, trends in the industry and the current status of the sales pipeline. That said, these forecasts are best treated as a baseline to work on rather than a firm prediction and can be used to anticipate weekly, monthly, quarterly or yearly sales revenue.

Sales forecasts help sales teams achieve their goals by identifying early warning signals in their sales pipeline and course-correct before it's too late

Sales forecasting also helps businesses to estimate their costs and revenue accurately based on which they are able to predict their short-term and long-term performance.

Sales forecasts can be used to identify benchmarks and determine incremental impacts of new initiatives, plan resources in response to expected demand, and project future budgets.

ML algorithms can extract, process, and learn from massive amounts of sales data. Models can analyze sales activities and customer data at scale, generate deeper insights, and even take action on those insights automatically.

Table of Contents

	Page No.
Acknowledgement	**V**
Abstract	**VI**
1. Introduction	**1**
1.1 Aims and Objectives	2
1.2 Research Questions	2
1.3 Background	3
1.3.1 Data Mining	3
1.3.2 Machine Learning	3
1.3.3 Machine Learning Algorithms	4
1.3.4 Selection of Machine Learning Algorithms	7
2. Literature Review	**8**
3. Methodology	**10**
3.1 Hypothesis Generation	10
3.2 Dataset Information	11
3.3 Objective	12
4. Flow Diagram	**13**
4.1 Experimental Environment	14
4.2 Data Overview	15
4.3 Feature Selection	16
4.4 Feature Importance	18
4.5 Data Preprocessing	18
5. Results	**18**
5.1 Exploratory Data Analysis	19
5.2 Model Development	26
6. Conclusion	**31**
7. Future Scope	**32**
8. References	**33**

List of Figures

		Page. No.
Figure 1.3.2	**Types of Machine Learning**	**3**
Figure 1.3.3 (a)	**Linear Regression**	**4**
Figure 1.3.3 (b)	**Decision Tree**	**6**
Figure 1.3.3 (c)	**Random Forest**	**6**
Figure 3.1	**Factors affecting sales**	**10**
Figure 5.1	**Univariate Analysis**	
	(a) Distribution of the variable Outlet_Size	**19**
	(b) Distribution of the variable Outlet_Location_Type	**20**
	(c) Distribution of the variable Item_Type	**21**
	(d) Distribution of the variable Outlet_Type	**21**
Figure 5.1	**Bivariate Distribution**	
	(a) Item_Weight and Item_Outlet_Sales analysis	**22**
	(b) Item_Visibility and Item_Outlet_Sales analysis	**23**
Figure 5.2	**Model Building**	
	(a) Linear Regression	**26**
	(b) Ridge Regression	**27**
	(c) Decision Tree Model	**28**
	(d) Random Forest Model	**29**

CHAPTER 1

INTRODUCTION

Predicting future sales for a company is one of the most important aspects of strategic planning. Our goal is to identify the most important variables and to define the best regression model for predicting our target variable.

Sales forecasting has always been a very significant area to concentrate upon. An efficient and optimal way of forecasting has become essential for all the vendors in order to sustain the efficacy of the marketing organizations. Manual infestation of this task could lead to drastic errors leading to poor management of the organization, and most importantly would be time consuming, which is something not desirable in this expedited world. A major part of the global economy relies upon the business sectors, which are literally expected to produce appropriate quantities of products to meet the overall needs.

Machine learning is the domain where the machines gain the ability to outperform humans in specific tasks. They are used to do some specialized task in a logical way and gain better results for the progress of the current society. The base of machine learning is the art of mathematics, with the help of which various paradigms can be formulated to approach the optimum output. In case of sales forecasting also machine learning has proved to be a boon. It is helpful in predicting the future sales more accurately.

BigMart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business. Machine learning can help us discover the factors that influence sales in a retail store and estimate the number of sales it will have in the near future.

1.1 Aims and Objectives

This thesis aims to develop a Machine Learning model that can predict the sales of products from different outlets. Several objectives were drawn to attain the goal.

Objectives :

- Converting data into an appropriate form using various preprocessing techniques for the implementation of Machine Learning algorithms.
- Finding critical features that will most influence sales of the product.
- To determine the appropriate Machine Learning algorithm for sales forecasting.
- Selecting various metrics to compare the performance of the applied Machine Learning algorithms.

1.2 Research Questions

Two research questions have been defined for this study to accomplish the aim. They are defined as follows:

RQ1:

What are the critical features that influence product sales?

Motivation:

The motivation of this research question is to find critical features in the data that can be useful while experimenting for RQ2 to build the Machine Learning model. This will help us reduce computational power and improves the quality of the results.

RQ2:

What is the best suitable algorithm for sales and demand prediction using Machine Learning techniques?

Motivation:

The critical features identified from RQ1 are used to develop the Machine Learning model using different algorithms. These models are compared by using various metrics such as accuracy score, mean absolute error, and max error to select the best fit model for the data.

1.3 Background

There are several methods for forecasting future demand for the goods and services a business provides. The forecasts are used for planning production and business activities, purchasing materials, inventory management, scheduling work hours, advertising, and often more across most industries. Traditional forecasting approaches were primarily focused on experienced employee opinions or statistical analysis of past data, but in recent years Machine Learning techniques have been implemented with great success in this field.

1.3.1 Data Mining

Data mining is described as a process for extracting usable data from a larger collection of raw data using statistical, artificial intelligence, Machine Learning and pattern recognition methods. Data Mining is increasingly seen as a step in a systematic and iterative process of knowledge discovery, in which automated pattern recognition methods are combined with expert knowledge of the analyst. This process is called the Knowledge Discovery in Databases (KDD) process.

1.3.2 Machine Learning

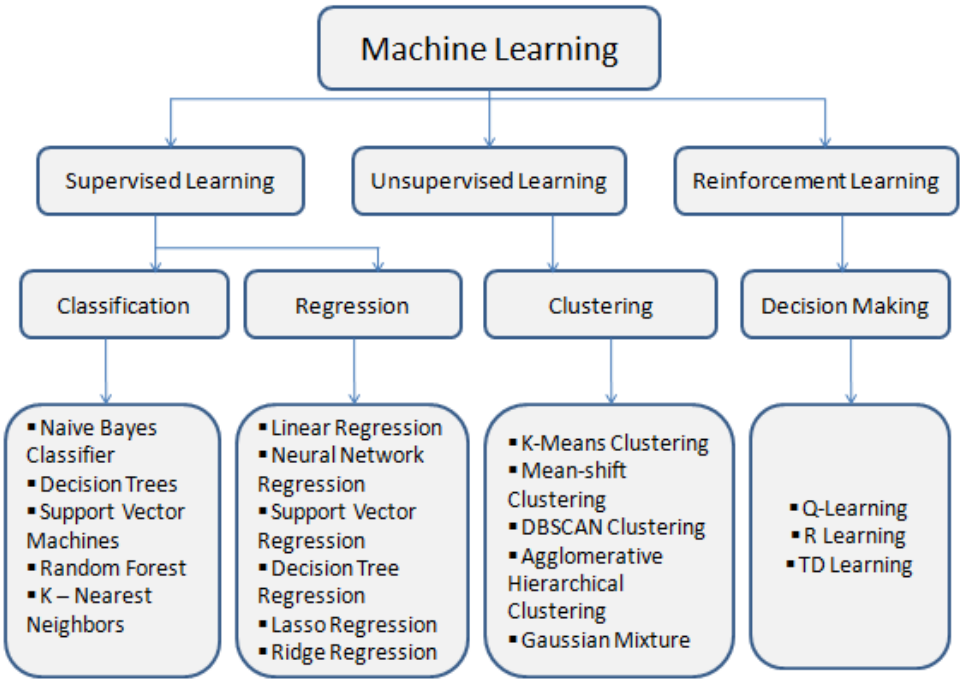


Figure 1.3.2 *Types of Machine Learning*

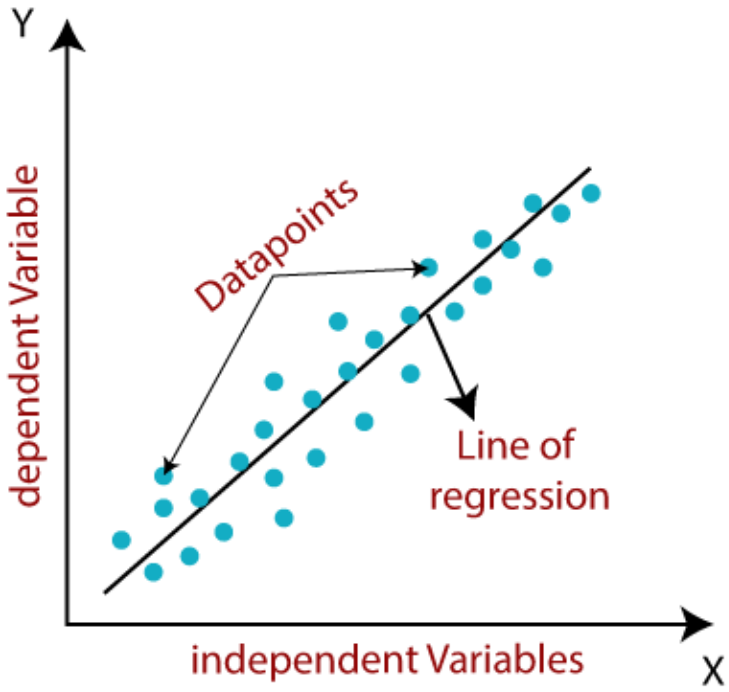
Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

1.3.3 Machine Learning Algorithms

Linear Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.



Mathematically, we can represent a linear regression as:

$$Y = a_o + a_1 x + \varepsilon$$

Ridge regression

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

The cost function for ridge regression:

$$Min(\left|\left|Y - X(\theta)\right|\right|^2 + \lambda||\theta||^2)$$

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced.

It shrinks the parameters. Therefore, it is used to prevent multicollinearity. It reduces the model complexity by coefficient shrinkage.

Ridge Regression Models -

For any type of regression machine learning model, the usual regression equation forms the base which is written as:

$$Y = eBX + e$$

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

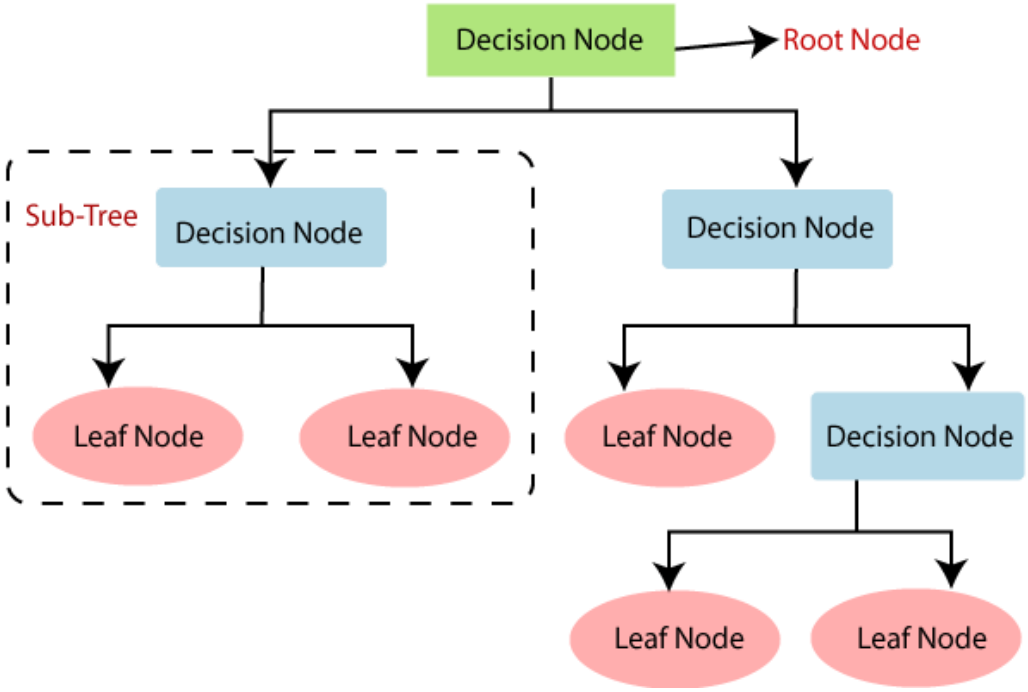
Once we add the lambda function to this equation, the variance that is not evaluated by the general model is considered. After the data is ready and identified to be part of L2 regularization, there are steps that one can undertake.

Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

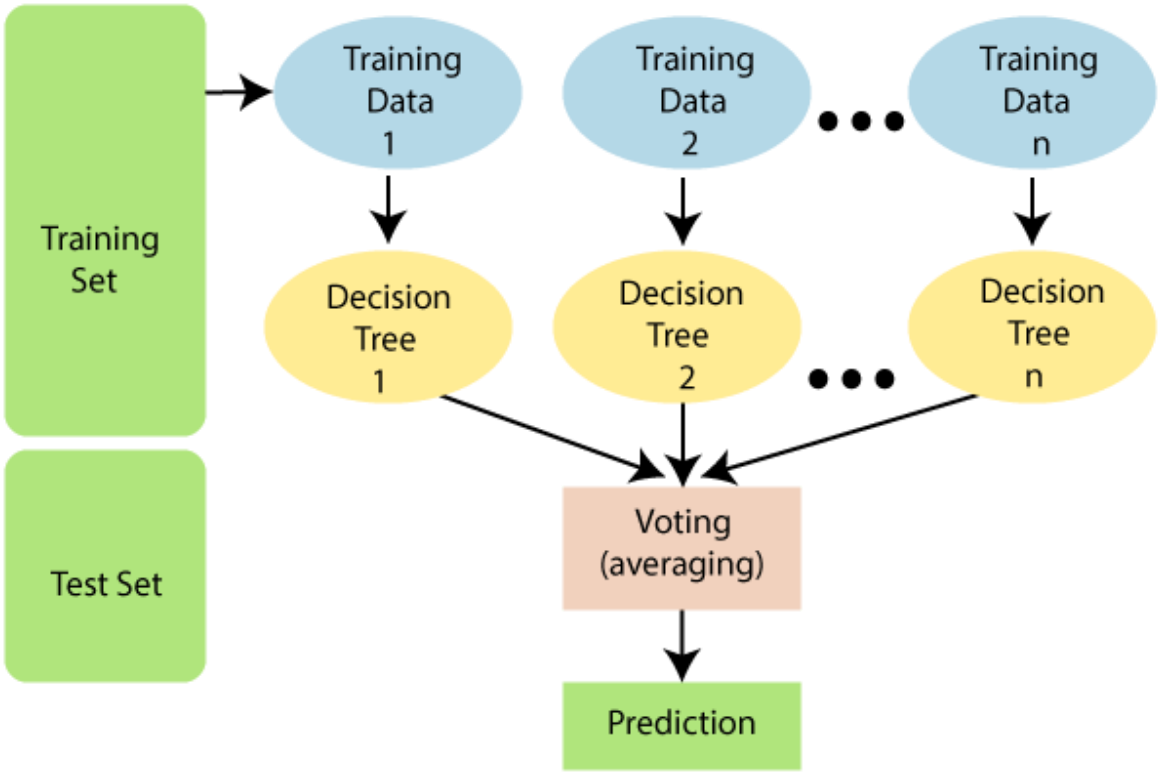
In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:



Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.



XGBoost

XGBoost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models.

The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees.

1.3.4 Selection of Machine Learning Algorithms

For every problem, choosing an algorithm is not a trivial decision. There is no proper algorithm that works for any problem, but few algorithms are widely recognized for performing the algorithms better than others in some cases. One can not assume the more accuracy from the algorithms for all types of data, accuracy will differ from data to data. In this thesis Machine Learning Algorithms such as Simple Linear Regression, Gradient Boosting Regression, Support Vector Regression, and Random Forest Regression were considered in which they expected to perform well on the issues.

CHAPTER 2

LITERATURE REVIEW

Previously a lot of sales and demand forecasting work was performed using Machine Learning. Most of the work in this research will concentrate on the sales of food items.

Due to the importance of forecasting in various fields, there are so many different types of approaches taken previously, some of the methods such as Machine Learning models, hybrid models, and statistical models. To handle this work, some of the statistical methods such as auto regressive moving average (ARMA) and auto regressive integrated moving average (ARIMA) will be helpful.

İrem İşlek and Şule Gündüz Öğüdücü experimented with the use of bipartisan graphic clusters that clustered different warehouses according to the sales behavior. They addressed the application by applying the Bayesian network algorithm in which they managed to produce the enhanced forecasting experience.

Grigorios Tsoumakas had used Machine Learning techniques to perform a survey on the forecasting of food sales. They had addressed data analyst design decisions such as temporal granularity, output variable, and input variables in this survey. In this paper the authors experimented by taking the point of sale (POS) as internal data and even external data by considering different environments to enhance the efficiency of demand forecasting. They considered different Machine Learning algorithms such as Boosted Decision Tree Regression, Bayesian Linear Regression, and Decision Forest Regression for evaluation.

The paper's authors had researched interestingly about customers coming to the restaurants using Random Forests, k-nearest neighbor, and XGBoost. They chose two real-world data sets from different booking sites and also made different input variables from restaurant features. The results have shown that XGBoost is the most appropriate model for the dataset. Holmberg and Halldén had observed that regular restaurant sales to be influenced by the weather. They considered two Machine Learning algorithms as XGBoost and neural network, and the results showed that the XGBoost algorithm is more accurate than the other

algorithm, and they also found that they had improved their model performance by 2-4 percentage points by taking weather factors into consideration. To improve accuracy, they had considered numerous variables such as characteristics, sales history, and weather factors.

Most of the recent studies focused on sales modeling without considering the relationship between the training and testing data, they used training data directly. This causes many errors which lead to a reduction in accuracy. Recent studies have suggested clustering techniques to separate the entire forecasting data into several clusters of predictable data before designing predictable models to minimize computational time and achieve effective evaluating performance.

In particular, Support Vector Machine(SVM) had been applied to demand forecasting. Garcia et al. (2012), in their study, proposed an intelligent model that relies on supporting vector machines to deal with issues relating to the allocation and revelation of new models. Kandananond (2012) showed that SVM surpassed Artificial Neural Networks in estimating demand for consumer goods.

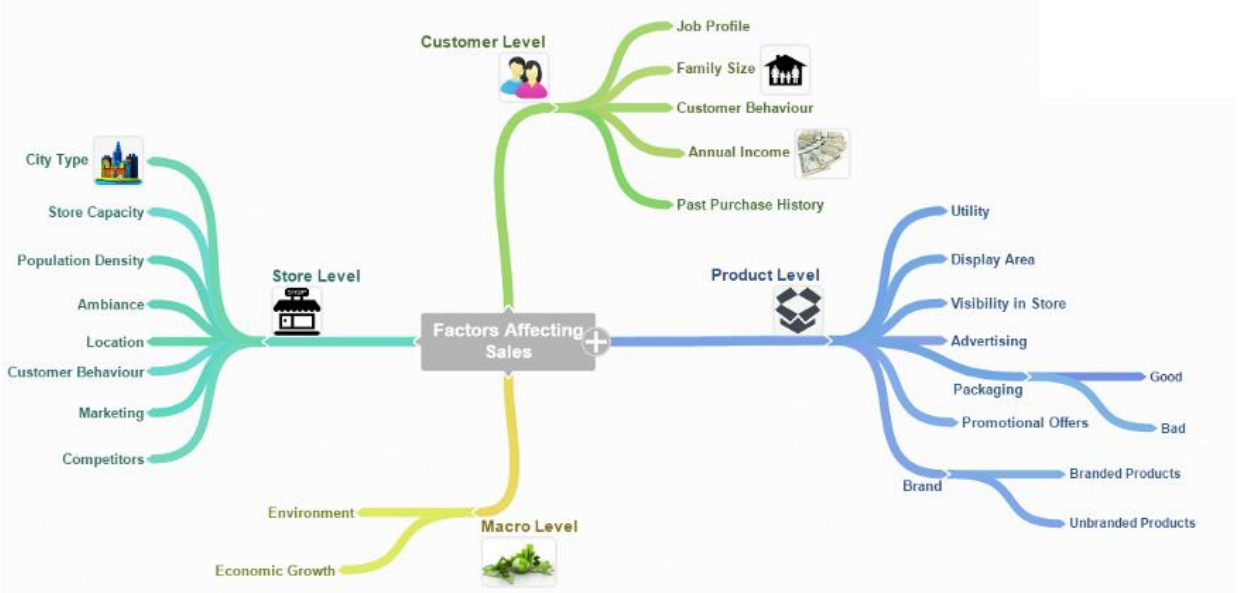
Previously, most of the studies focused on considering the metrics as mean absolute error, mean squared error, median absolute error, and k-fold cross validation is used for training and testing data. Metrics like max error, accuracy, and mean absolute error are considered in this research. In this study stratified K-fold crossvalidation technique is used for training and testing to increase the efficiency of the results. In this study a suitable algorithm is chosen for sales forecasting.

CHAPTER 3

METHODOLOGY

3.1 Hypothesis Generation

It is always a good idea to generate some hypothesis before proceeding to any data science project. We can separate this process into four levels: Product level, Store level, Customer level, and Macro level.



Product level hypothesis :

- Brand : Branded products have more trust of the customers so they should have high sales.
- Visibility in Store: The location of the product placement also depends on the sales
- Display Area: Products that are placed at an attention-catching place should have more sales.
- Utility: Daily use products have a higher tendency to sell compared to other products.
- Packaging: Quality packaging can attract customers and sell more.

Store Level Hypothesis :

- City type: Stores located in urban cities should have higher sales.
- Store Capacity: One-stop shops are big in size so their sell should be high.
- Population density: Densely populated areas have high demands so the store located in these areas should have higher sales.
- Marketing: Stores having a good marketing division can attract customers through the right offers.

3.2 Dataset Information

The dataset we will use in this project is downloaded from Kaggle. The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

Link : https://www.kaggle.com/devashish0507/big-mart-sales-prediction

Name	Type	Subtype	Description	Segment	Expectation
Item_Identifier	Numeric	Discrete	Unique Product ID	Product	Low Impact
Item_weight	Numeric	Continuous	Weight of product	Product	Medium Impact
Item_Fat_Content	Categorical	Ordinal	Wether the product is low fat or not	Product	Medium Impact
Item_Visibility	Numeric	Continuous	% of total display area in store allocated to this product	Product	High Impact
Item_Type	Categorical	Nominal	Category to which product belongs	Product	High Impact
Item_MRP	Numeric	Discrete	Maximum Retail Price (list price) of product	Product	Medium Impact
Outlet_Identifier	Numeric	Discrete	Unique Store ID	Store	Low Impact
Outlet_Establishment_Year	Numeric	Discrete	Year in which store was established	Store	Low Impact
Outlet_Size	Categorical	Ordinal	Size of the store	Store	High Impact
Outlet_Location_Type	Categorical	Ordinal	Type of city in which the store is located	Store	High Impact
Outlet_Type	Categorical	Ordinal	Grocery store or some sort of supermarket	Store	High Impact
Item_Outlet_Sales	Numeric	Discrete	Sales of product in particular store. This is the outcome variable to be predicted	Product	Target

3.3 Objective

The objective of our project is to develop a statistical Model based on the dataset available.

- Step 1 - First we will import modules and load data using pandas.
- Step 2 - We will then preprocess the big mart sales data and create new attributes.
- Step 3 - Exploratory Data Analysis of big mart sales
- Step 4 - Feature Engineering
- Step 5 - We are going to use different models to test the accuracy and will finally train the whole data to check the score.

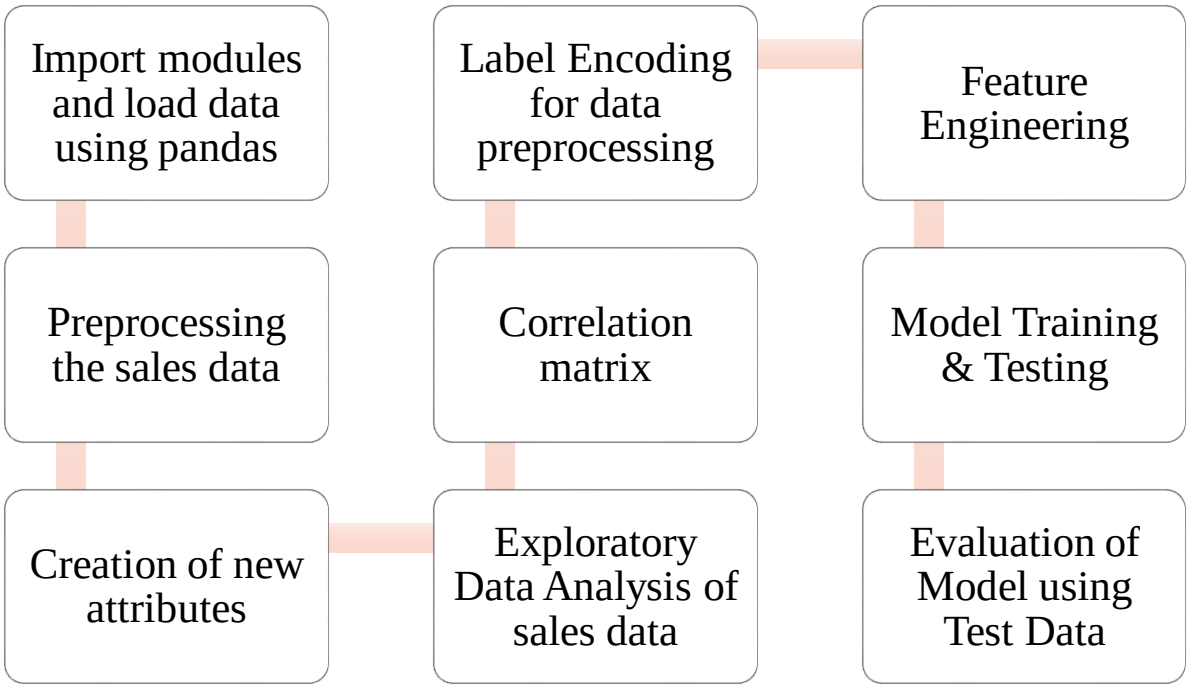
CHAPTER 4

FLOW DIAGRAM

We used pandas library to analyze the data. Matplotlib & Seaborn to visualize the data and scikit learn for statistical modeling. First we import the modules and load our dataset using pandas library. Then we will preprocess the data.

Data preprocessing is divided into four stages: data cleaning, data integration, data reduction, and data transformation. Data preprocessing has the objective to add missing values, aggregate information and label data with categories (Data binning) .

We removed the outliers and created some new attributes. The next step is to understand and visualize the data. We used histograms, density and scatterplot for visualizing our data. We then created a correlation matrix. A correlation matrix is a basically a table which shows correlation coefficients between variables.



Each cell in the table shows the correlation between two variables. Then we performd label encoding. Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. For this We used the scikit-learn library approach. We first Created an instance of LabelEncoder() and stored it in labelencoder variable we then Applied fit and transform method to assign numerical value to categorical value. Then we performed Feature Engineering. Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used when creating a predictive model using machine learning or statistical modeling.

4.1 Experimentation Environment

An experiment is chosen for the first research question i.e. correlation. Each data attribute can be selected by applying feature selection methods like data correlation and which will make the predictable attributes more accurate. This will reduce a lot of strain on the Machine Learning model during pre-processing and cleansing the data. For the second research question an experiment is chosen because the experiments provide control over factors and a deeper understanding of many common research techniques such as a case study or survey. One can describe the procedure followed in this experiment as follows:

- Extracting the data required for the sales.
- Applying specified Machine Learning (supervised) algorithms.
- The performance of the output can be enhanced by comparing metrics such as accuracy score, mean absolute error and max error.
- Based on assessment tests, the best suitable algorithm can be selected.

Python

Python is a commonly used high-level programming language, it was designed by Guido van Rossum which can be easy to interpret and read. Python has specific functionality and is convenient to be used for both quantitative and analytical computational purposes. Data Science Python is popularly used and, as well as being a dynamic and open source language, is a top choice. Its massive libraries are also used to manipulate the data however for a beginner data analyst they are really simple to learn. The python libraries used in this thesis are briefly described as follows :

1. **Pandas**

Pandas is a software library that is designed for manipulating the data and analysis in a python programming language. It is open-source which is released under the BSD license of three clauses. It is based on the Numpy package, and the DataFrame is its main data structure.

2. **Matplotlib**

Matplotlib is a module of Python used to plot the attractive Graphs. Visual representation in data science is a significant step. One can quickly understand how data is split by using visual representation. There are many libraries to represent the data, but the matplotlib is very widely known and easier to visualize.

3. **Sklearn**

Scikit-learn is a free python library. It features multiple clustering classification and regression algorithms including random forests, DBSCAN, k-means, gradient boosting, support vector machines, and gradient boosting which is programmed to interface with the NumPy and SciPy libraries.

4. **Seaborn**

Seaborn is a open-source python library that is used for statistical graphics. It offers a data set-oriented API to analyze relationships among different variables, as well as resources to select color palettes that truly in the data.

4.2 Data overview

In this thesis, there is labeled sales data from different items from different outlets that provide information such as item type, item price, outlet type, etc. These data were extracted from various sources and will be used to train and improve the model for Machine Learning. In the dataset being analyzed there are 8523 instances and 12 attributes. The dataset has been properly divided into training and testing data that can be described in the sections below.

4.3 Feature Selection

There are various types of factors that can make the model of Machine Learning more effective on any given task. One of the methods of feature selection is data correlation which will have a major impact on the model's performance. This will reduce a lot of strain on the Machine Learning model during preprocessing and cleansing the data.

The data attributes chosen for training the Machine Learning model would have a major impact on the efficiency of the model. Because of the irrelevant features that are presented, the model output will be reduced. The feature selection method provides an efficient way to remove data redundancy and irrelevant data that helps to reduce computation time, improve accuracy, and also enhance understanding of the model.

	Item Weight	*Item Visibility*	*Item MRP*	*Outlet Establishment Year*	*Item Outlet Sales*
count	7060.000	8523.000000	8523.0000	8523.000000	8523.000000
mean	12.857645	0.066132	140.99278	1997.831867	2181.288914
Std	4.643456	0.051598	62.275067	8.371760	1706.499616
Min	4.55500	0.000000	31.290000	1985.00000	33.290000
25%	8.773750	0.026989	93.82650	1987.000000	834.247400
50%	12.600000	0.053931	143.01280	1999.00000	1794.331000
75%	16.850000	0.094585	185.64370	2004.000000	3101.29640
Max	21.35000	0.328391	266.88840	2009.000000	13086.964800

The selection of features plays a crucial role in classification and involves selecting a subset of features that reflect the complete attributes that currently exist. Feature selection techniques are intended to improve classification efficiency by selecting the essential features from the data sets according to particular algorithms.

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
Item_Weight	1.000000	-0.014048	0.027141	-0.011588	0.014123
Item_Visibility	-0.014048	1.000000	-0.001315	-0.074834	-0.128625
Item_MRP	0.027141	-0.001315	1.000000	0.005020	0.567574
Outlet_Establishment_Year	-0.011588	-0.074834	0.005020	1.000000	-0.049135
Item_Outlet_Sales	0.014123	-0.128625	0.567574	-0.049135	1.000000

Data Columns	Number of non-null values
Item Identifier	8523 non-null objects
Item Weight	7060 non-null float64
Item Fat Content	8523 non-null objects
Item Visibility	8523 non-null float64
Item Type	8523 non-null objects
Item MRP	8523 non-null float64
Outlet Identifier	8523 non-null objects
Outlet Establishment Year	8523 non-null int64
Outlet Size	6113 non-null objects
Outlet Location Type	8523 non-null objects
Outlet Type	8523 non-null objects
Item Outlet Sales	8523 non-null float64

4.4 Feature Importance

Feature Importance refers to a class of approaches for assigning values to input features to a predictive model which determines the relative significance of each factor while forecasting.

Feature importance scores provide overview into the model. Most significant scores are determined using a prediction approach that was fitted to the dataset. Inspecting the score of importance gives insight into that particular model and what features are the most essential and least important to the model while making a prediction. This is a type of interpretation of the model that can be carried out for those models that encourage it.

Feature Importance can be used to enhance a predictive model. This can be accomplished by selecting those features to remove (lowest scores) or those features to retain, using the importance scores. This is a type of selection of features, and can simplify the modelling problem, accelerate the modelling process, and in certain cases improve model performance.

4.5 Data preprocessing

Before applying Machine Learning algorithms some of the missing values have been found which can impact the model's output so this should be handled. The 'item weight' and 'outlet size' attributes have 17 percent, and there is 28 percent of missing values. To make the dataset more efficient, these missing values will be replaced by the most promising values. There's more correlation between two of the different attributes with similar work. Removing one of the attributes will make the work better. The redundant values such as LF and reg provided in the attribute of item fat content will be treated and these redundant values will be replaced accordingly. The least value for an 'item visibility' attribute is zero which makes no sense for the dataset.

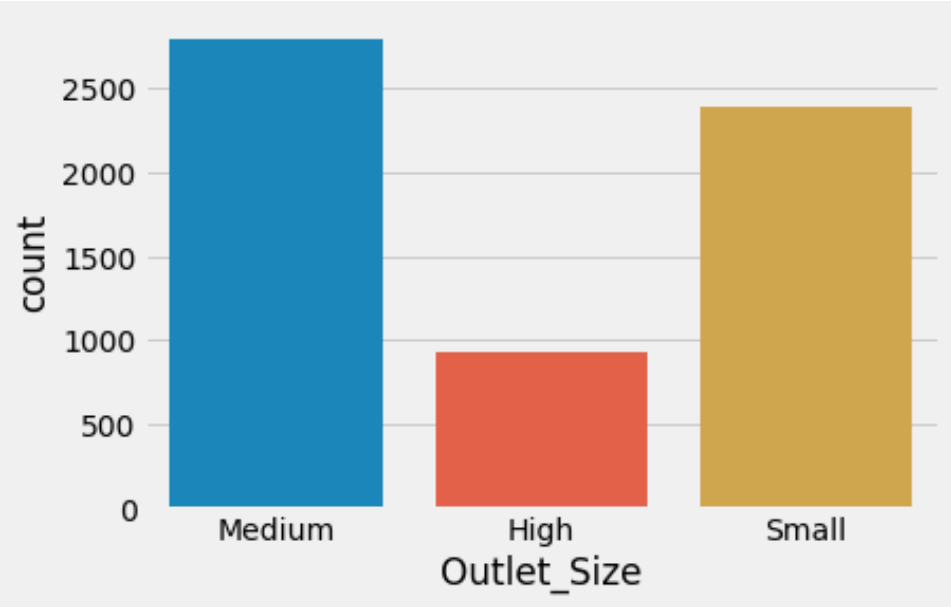
CHAPTER 5

RESULTS

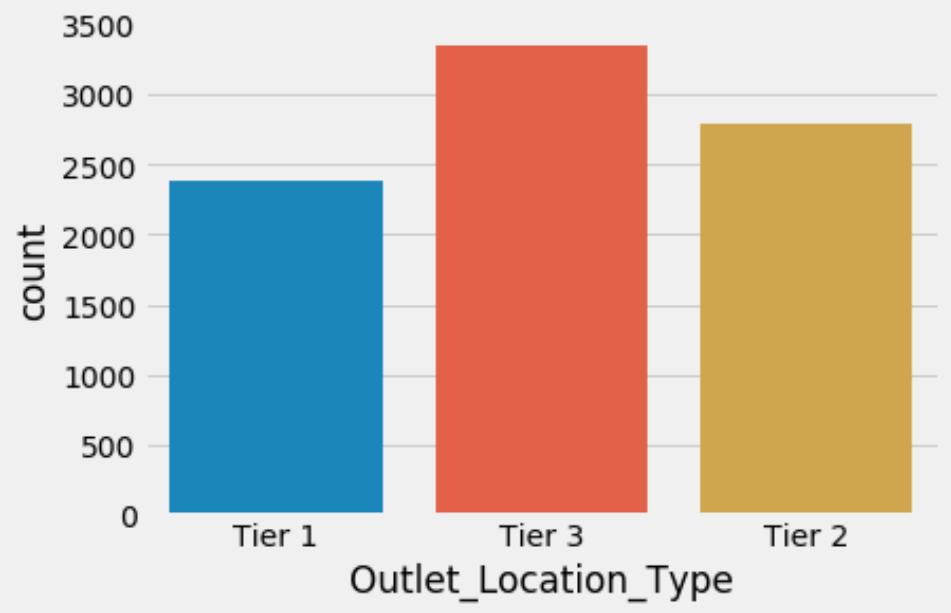
5.1 Exploratory Data Analysis

We've made our first assumptions on the data and now we are ready to perform some basic data exploration and come up with some inference. Hence, the goal for this section is to take a glimpse on the data as well as any irregularities so that we can correct on the next section, Data Pre-Processing.

Univariate Analysis –



(a) Distribution of the variable Outlet_Size



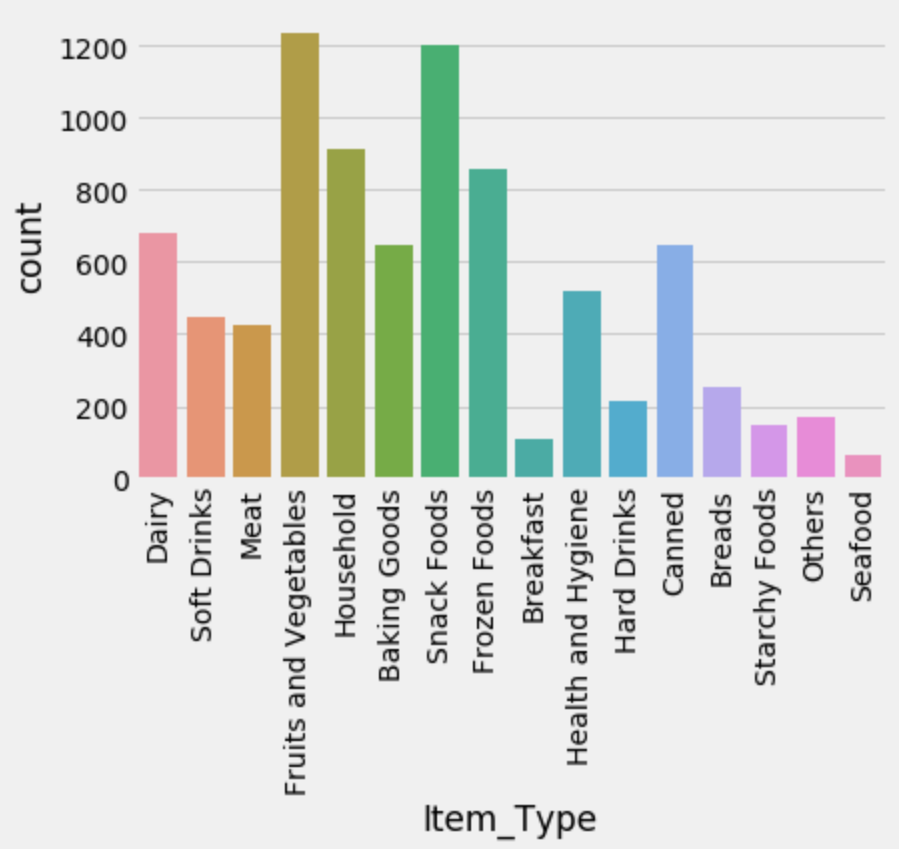
(b) Distribution of the variable Outlet_Location_Type

We first did univariate analysis. In univariate analysis there is only one dependable variable. We used histograms to get an idea of the distribution of numerical variables.

(a) There seems to be a low number of stores with size equals to “High”. Most of the existent stores seem to be either “Small” or “Medium”.

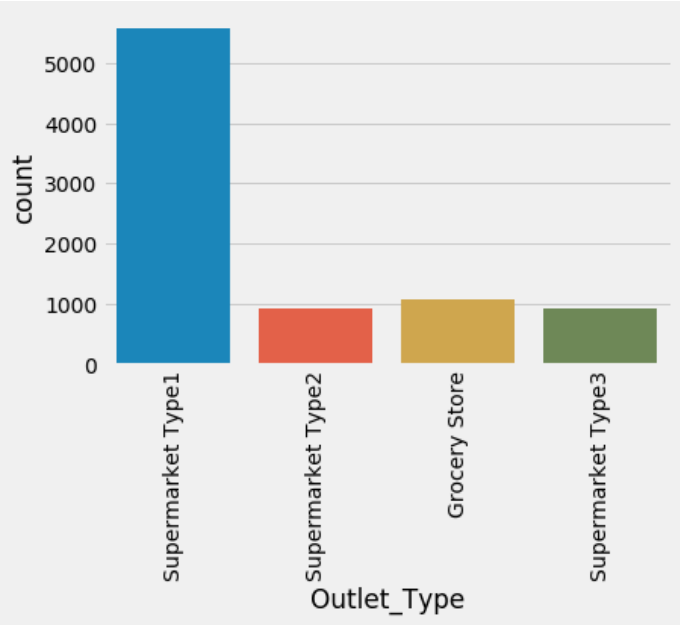
(b) Now in the second chart we can see that Bigmart appears to be a supermarket brand that is more present in “Small” to “Medium” size cities than in more densily populated locations.

(c) Looking at the list of Item_Type we see there are sixteen different types. This is a high number of unique values for a categorical variable. Most of them are fruits and vegetables while the least are seafood items.



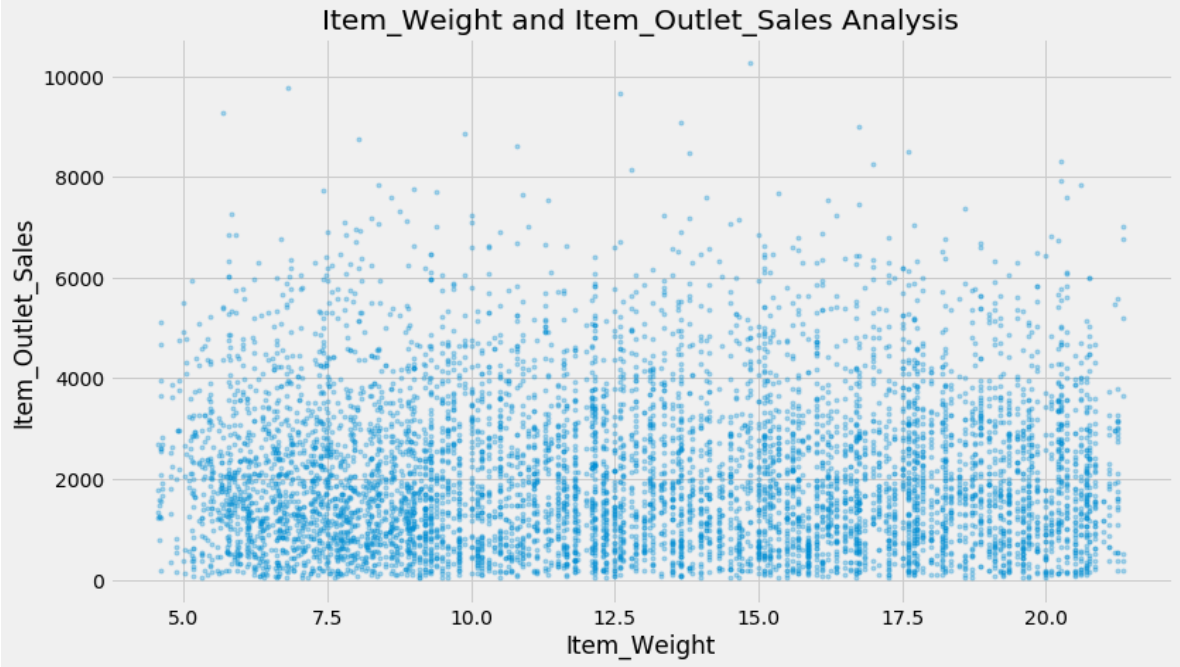
(c) Distribution of the variable Item_Type

(d) Here we can see that Supermarket Type2 , Grocery Store and Supermarket Type3 all have low expression in this distribution.

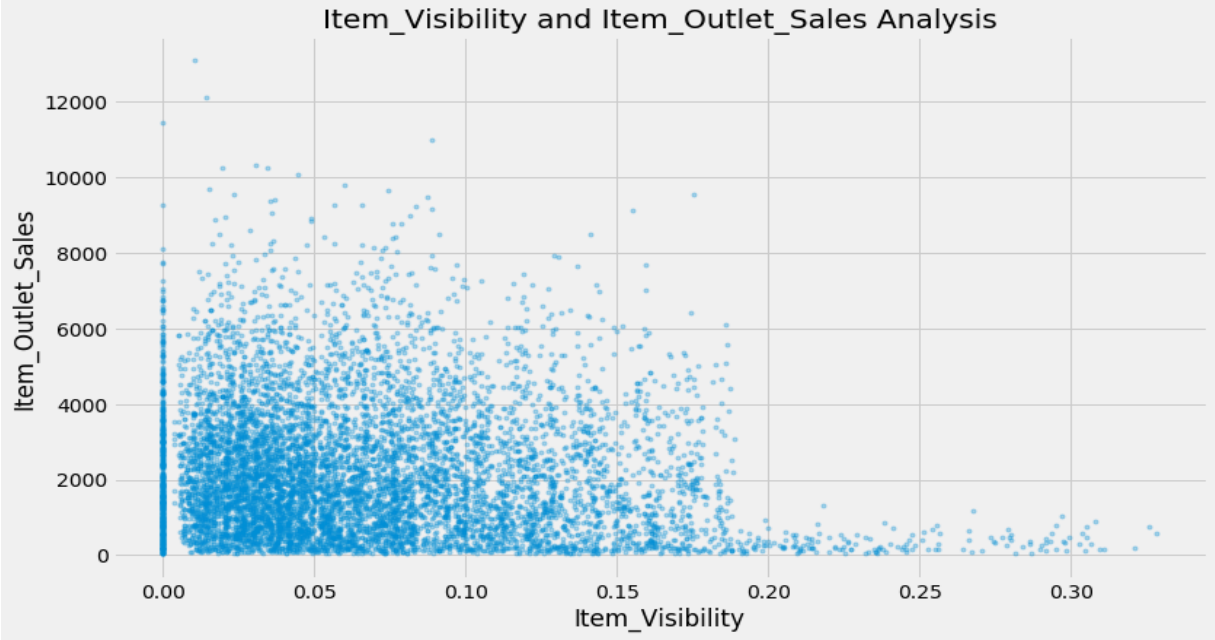


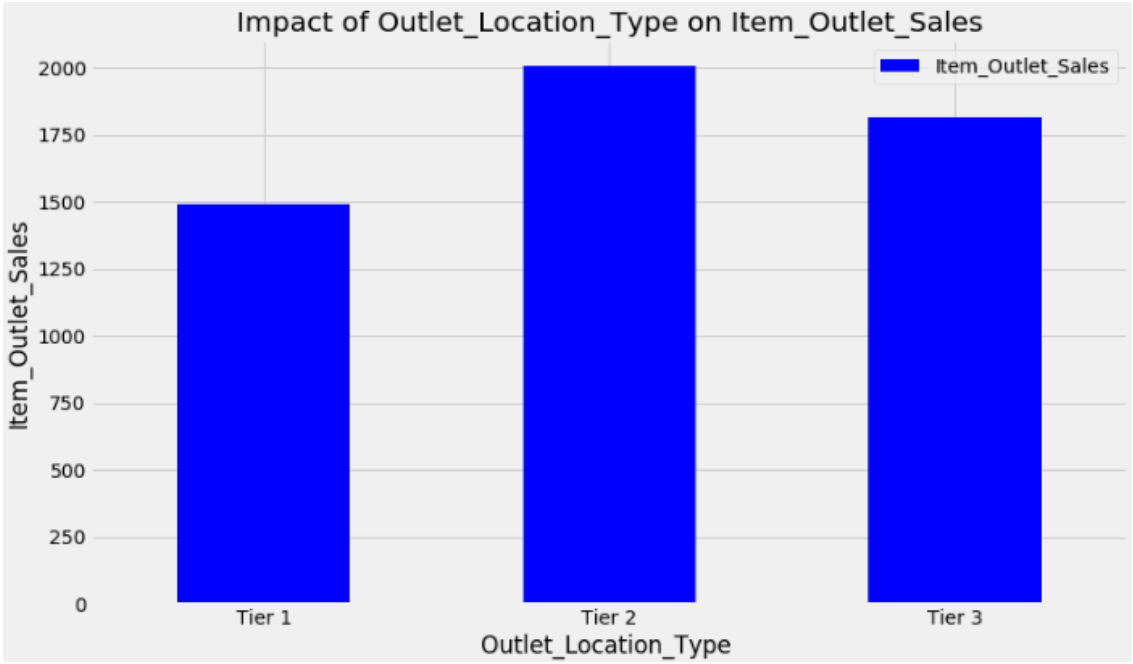
(d) Distribution of the variable Outlet_Type

Bivariate Distribution –



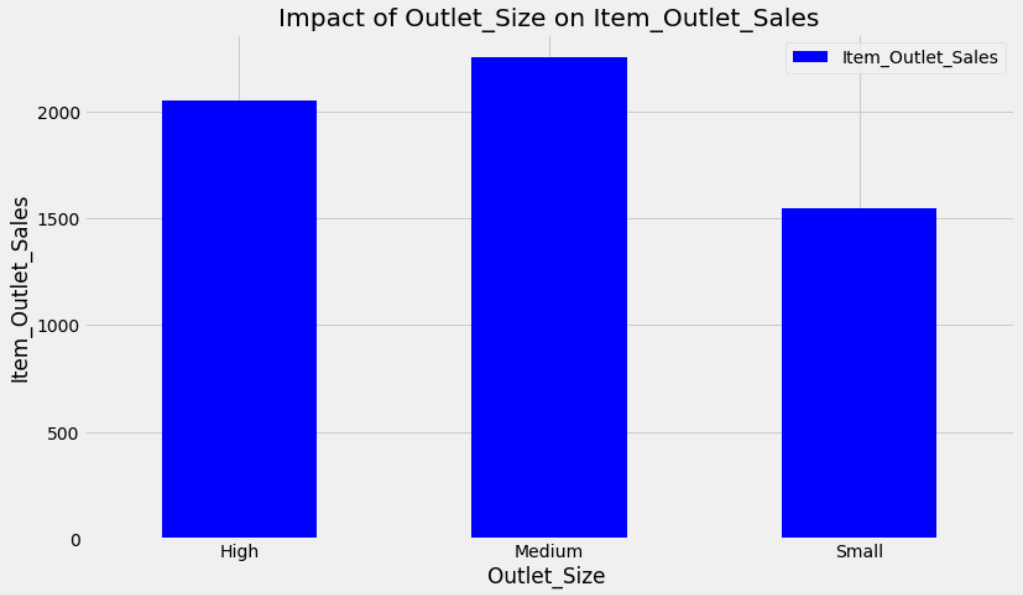
After univariate we performed bivariate analysis. It is a statistical test that involves two separate variables. It is used to determine whether or not two variables are related. In this plot chart we can see that Item_Weight has a low correlation with our target variable. So it will have little to no impact on our future analysis. The location of product in a store will impact sales. Generally the items which are right at entrance catches the eye of customer first rather than the ones in back. But according to this plot chart we discovered that the less visible the product is the more higher its sales will be. This might be due to the fact that a great number of daily use products, which do not need high visibility, control the top of the sales chart.



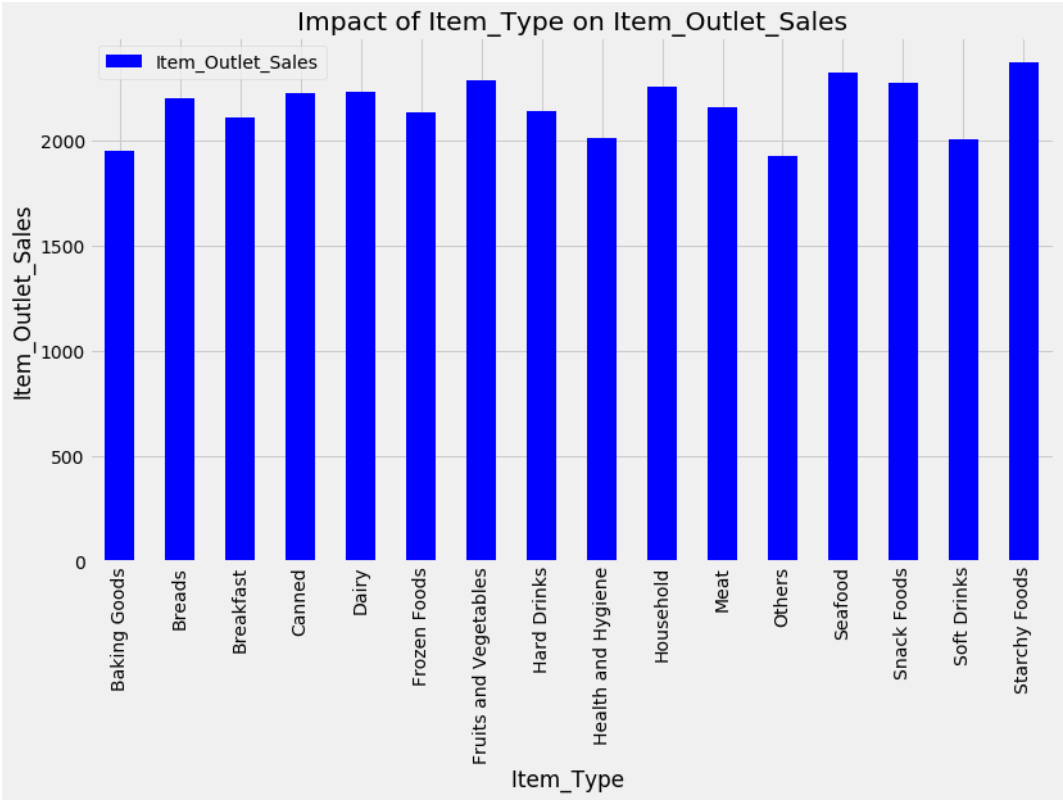


Do Tier 1 cities have higher sales?

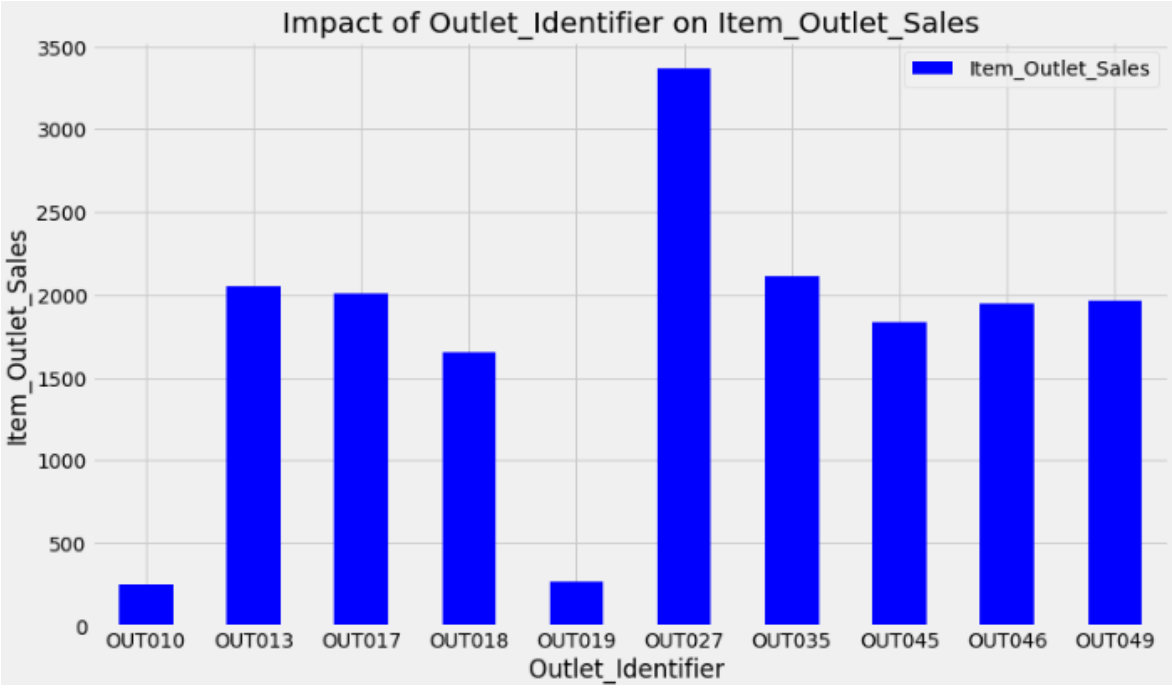
If we look at our results we see that stores from Tier 2 cities have highest product sales, followed by Tier 3 cities.



Our belief was that stores which are very big in size should have higher sales as they act like one-stop-shops and people would prefer getting everything from one place. But if we look at the results we can see that consumers tend to prefer medium size stores instead of big size.



Now let's look at that impact of item type on our target variable. Here we can see that Starchy food items have the highest sales while baking goods, health & hygiene items have the lowest sales.



Outlet_Identifier	OUT010	OUT013	OUT017	OUT018	OUT019	OUT027	OUT035	OUT045	OUT046	OUT049
Outlet_Type	Grocery Store	Supermarket Type1	Supermarket Type1	Supermarket Type2	Grocery Store	Supermarket Type3	Supermarket Type1	Supermarket Type1	Supermarket Type1	Supermarket Type1

Now let's look at the Impact of Outlet_Identifier on our target variable. From the ten stores, two are Groceries whereas six are Supermarket Type1, one Supermarket Type2 and one Supermarket Type3. We see that the groceries have the lowest sales results which is expected, followed by the Supermarket Type 2. Curiously, most stores are of type Supermarket Type1 of size "High" and do not have the best results. The best results belong to "Medium" size Supermarket Type 3.

5.2 Model Development

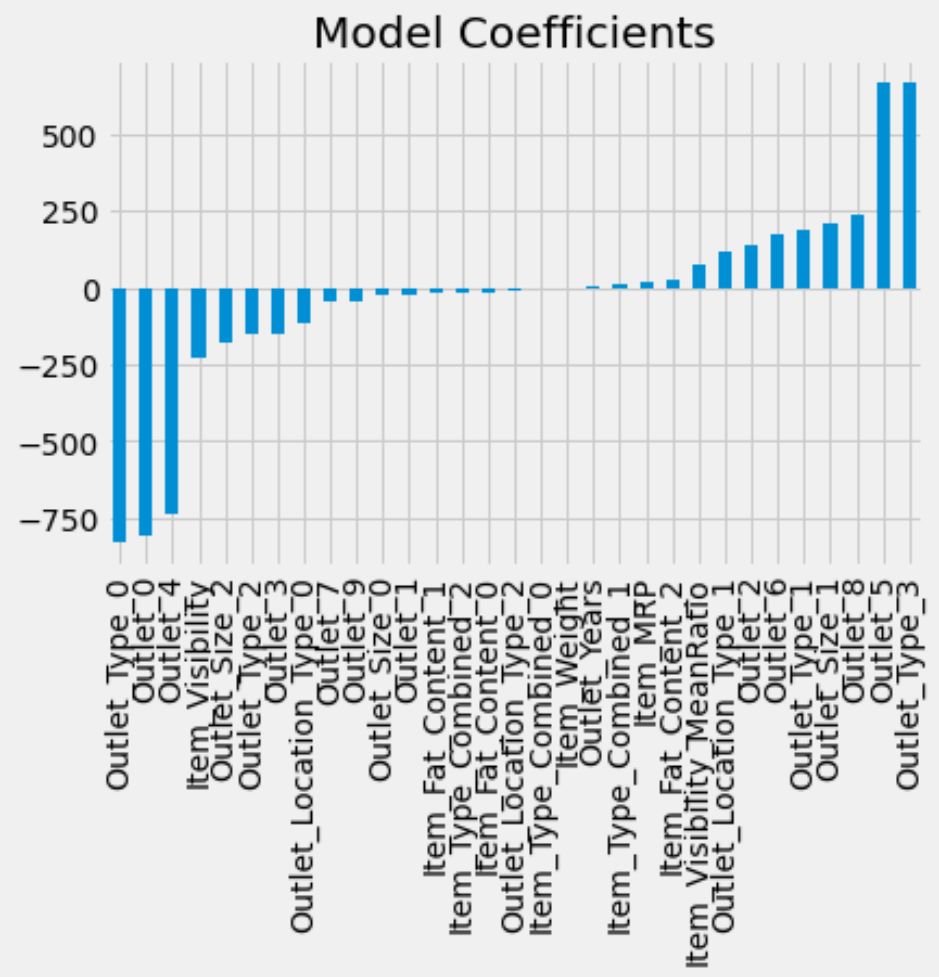
Linear Regression –

```
from sklearn.linear_model import LinearRegression
LR = LinearRegression(normalize=True)

predictors = train_df.columns.drop(['Item_Outlet_Sales','Item_Identifier','Outlet_Identifier'])
modelfit(LR, train_df, test_df, predictors, target, IDcol, 'LR.csv')

coef1 = pd.Series(LR.coef_, predictors).sort_values()
coef1.plot(kind='bar', title='Model Coefficients')
```

```
Model Report
RMSE : 1127
CV Score : Mean - 1129 | Std - 43.46 | Min - 1075 | Max - 1210
r2 score : 0.5635
```

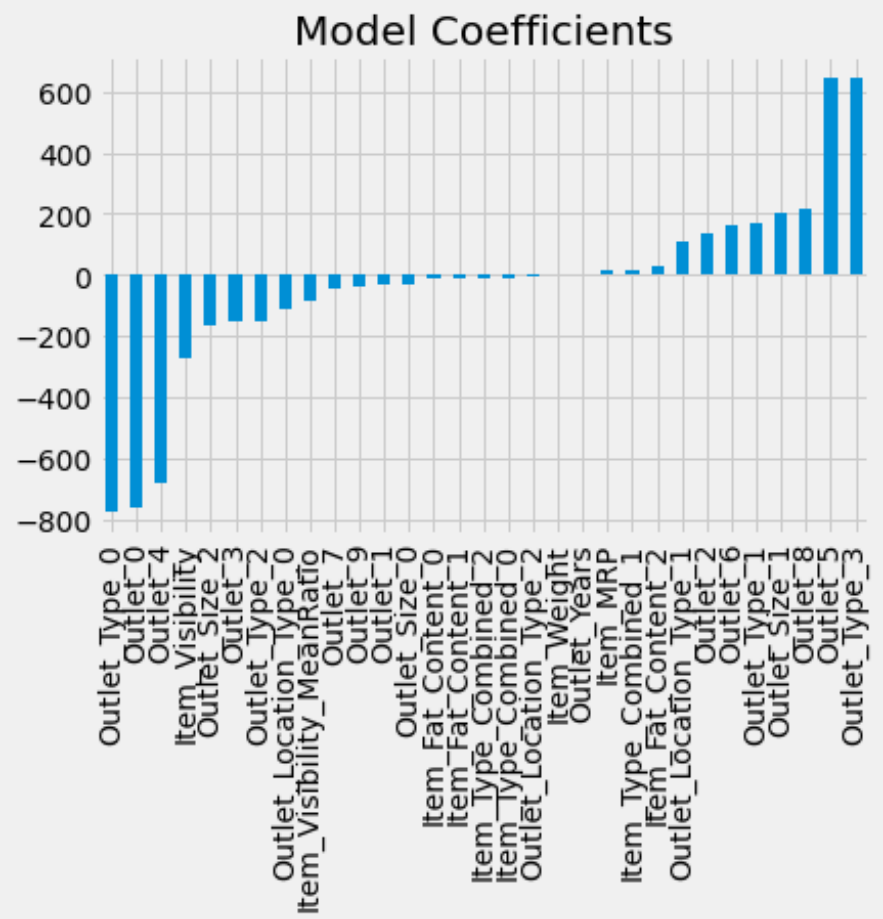


Ridge –

```
from sklearn.linear_model import Ridge
RR = Ridge(alpha=0.05,normalize=True)
modelfit(RR, train_df, test_df, predictors, target, IDcol, 'RR.csv')

coef2 = pd.Series(RR.coef_, predictors).sort_values()
coef2.plot(kind='bar', title='Model Coefficients')
```

```
Model Report
RMSE : 1129
CV Score : Mean - 1130 | Std - 44.6 | Min - 1076 | Max - 1217
r2 score : 0.5625
```

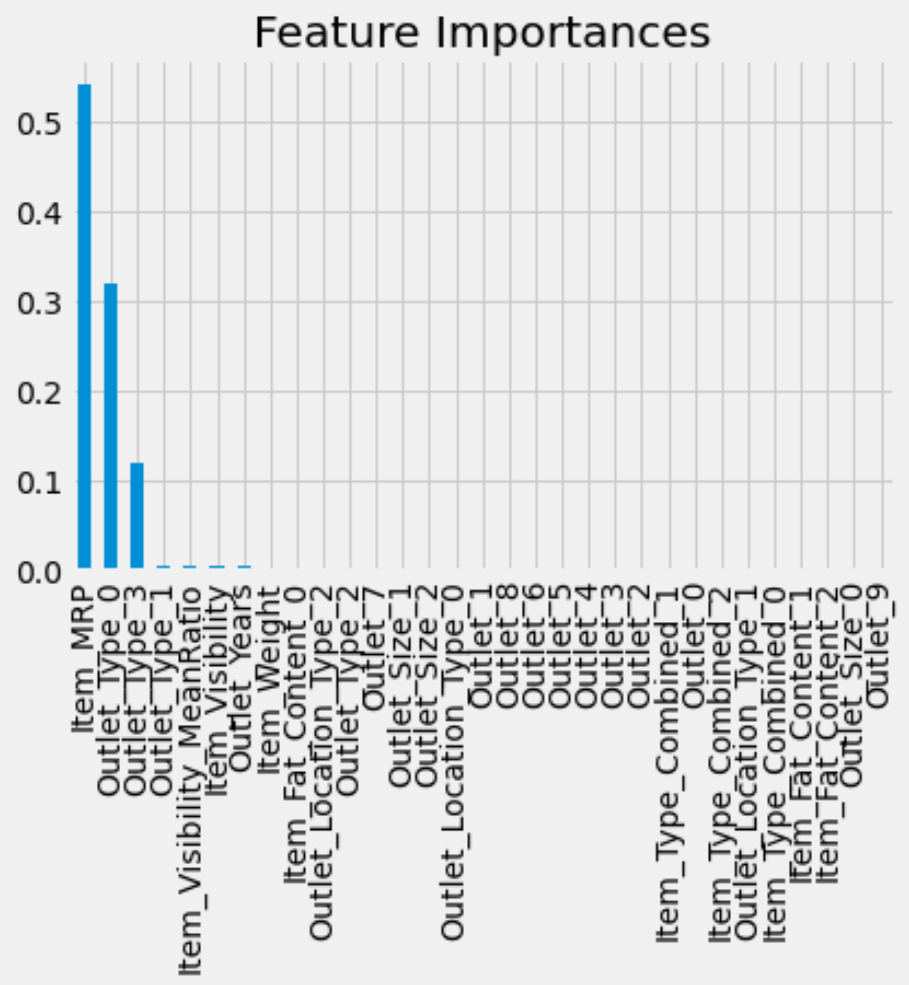


Decision Tree Model –

```
from sklearn.tree import DecisionTreeRegressor
DT = DecisionTreeRegressor(max_depth=15, min_samples_leaf=100)
modelfit(DT, train_df, test_df, predictors, target, IDcol, 'DT.csv')

coef3 = pd.Series(DT.feature_importances_, predictors).sort_values(ascending=False)
coef3.plot(kind='bar', title='Feature Importances')
```

```
Model Report
RMSE : 1058
CV Score : Mean - 1091 | Std - 45.42 | Min - 1003 | Max - 1186
r2 score : 0.6158
```

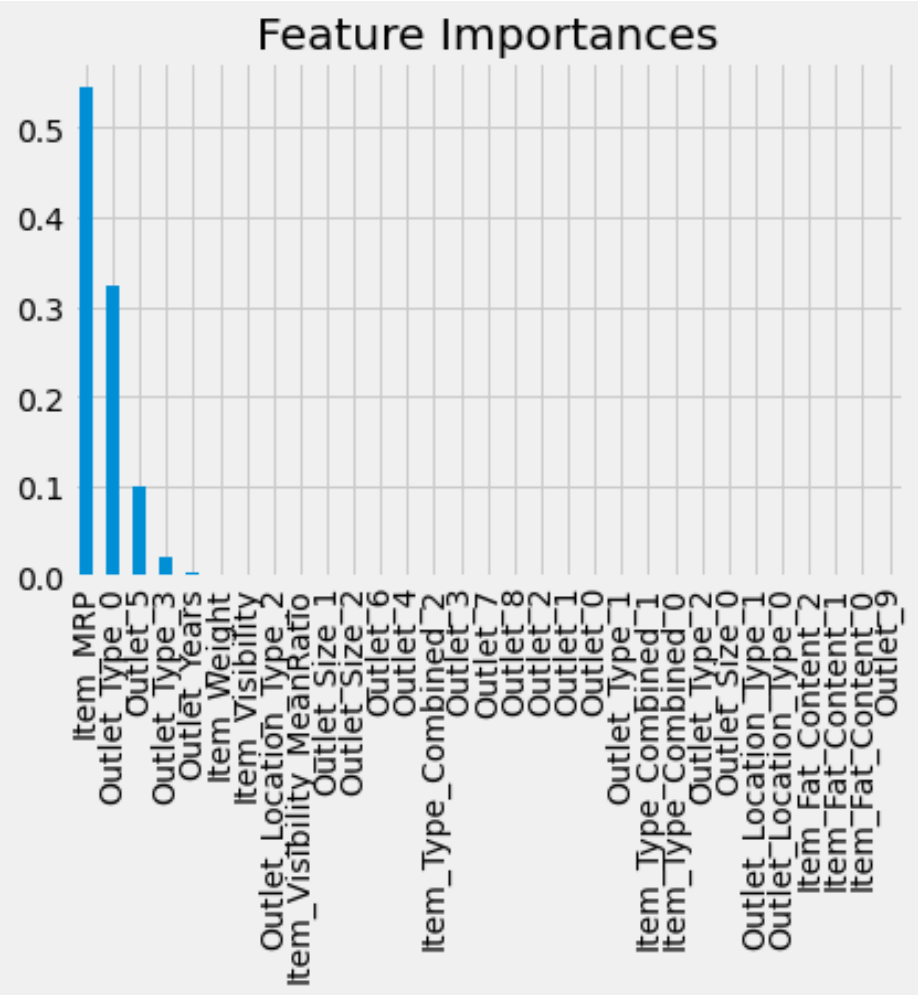


Random Forest Model –

```
RF = DecisionTreeRegressor(max_depth=8, min_samples_leaf=150)
modelfit(RF, train_df, test_df, predictors, target, IDcol, 'RF.csv')

coef4 = pd.Series(RF.feature_importances_, predictors).sort_values(ascending=False)
coef4.plot(kind='bar', title='Feature Importances')
```

```
Model Report
RMSE : 1069
CV Score : Mean - 1097 | Std - 43.41 | Min - 1028 | Max - 1180
r2 score : 0.6077
```



XgBoost –

```
from xgboost import XGBRegressor

my_model = XGBRegressor(n_estimators=1000, learning_rate=0.05)
my_model.fit(train_df[predictors], train_df[target], early_stopping_rounds=5,
             eval_set=[(test_df[predictors], test_df[target])], verbose=False)
```

```
XGBRegressor(learning_rate=0.05, n_estimators=1000)
```

```
#Predict training set:
train_df_predictions = my_model.predict(train_df[predictors])

# make predictions
predictions = my_model.predict(test_df[predictors])
```

```
Mean Absolute Error : 129.90780382232188
RMSE : 1052
r2 score : 0.6197
```

CHAPTER 6

CONCLUSION

Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, sales revenue analysis will help to get the details needed to estimate both the revenue and the income. Different types of Machine Learning techniques such as Ridge Regression, Decision Tree, Simple Linear Regression, Random Forest and XGBoost Model have been evaluated on sales data to find the critical factors that influence sales to provide a solution for forecasting sales. After performing metrics such as accuracy, mean absolute error, and max error, the XGBoost is found to be the appropriate algorithm with accuracy of 61.97% according to the collected data and thus fulfilling the aim of this thesis.

Algorithms	Accuracy
Linear Regression	56.35%
Ridge Regression	56.25%
Decision Tree	61.58%
Random Forest	60.77%
XGBoost Model	61.97%

CHAPTER 7

FUTURE SCOPE

With satisfactory effectively to amplify our answer to assist shops enhances productiveness and expands income by using taking benefit of information analysis. Sales prediction performs a necessary function in growing the effectively with which shops can function as it presents important points on the visitors a save can count on to get hold of on a given day. In addition to simply predicting the contemplated sales, there are different facts which can be mined to spotlight essential tendencies and additionally enhance planning. Such are advertisement, recommendation, predicting demand, consumer based totally pricing, holiday/extended sale planning and product classification.

In future work one can attempt performance metrics such as time while predicting the sales. These metrics can play a crucial role in evaluating multiple Machine Learning algorithms. And also one can attempt to implement more accurate data in the continued study. Machine Learning has the advantage of analyzing data and key variables so that you can aim to develop a systematic approach using a variety of Machine Learning techniques.

CHAPTER 8

REFERENCES

[1] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. Machine learning methods for demand estimation. American Economic Review, 105(5):481–85, 2015.

[2] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1):69–88, 2016.

[3] Ankur Jain, Manghat Nitish Menon, and Saurabh Chandra. Sales forecasting for retail chains, 2015.

[4] Grigorios Tsoumakas. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1):441–447, 2019.

[5] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3):275–294, 2012.

[6] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. Journal of the american statistical association, 83(404):1023–1032, 1988.

[7] Zheng Li, Xianfeng Ma, and Hongliang Xin. Feature engineering of machinelearning chemisorption models for catalyst design. Catalysis today, 280:232–238, 2017.

[8] Xinchuan Zeng and Tony R Martinez. Distribution-balanced stratified crossvalidation for accuracy estimation. Journal of Experimental & Theoretical Artificial Intelligence, 12(1):1–12, 2000.

[9] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 145–158. Springer, 2011.

[10] Chris Rygielski, Jyun-Cheng Wang, and David C Yen. Data mining techniques for customer relationship management. Technology in society, 24(4):483–502, 2002

[11] Krzysztof J Cios, Witold Pedrycz, Roman W Swiniarski, and Lukasz Andrzej Kurgan. Data mining: a knowledge discovery approach. Springer Science & Business Media, 2007.