DEVELOPMENT OF AN APPLICATION SOFTWARE FOR SALES PREDICTION USING MACHINE LEARNING ALGORITHMS

FERJANA AHMED (SN. 1017140016)

A Thesis Submitted in Partial Fulfilment of the Requirements for The Degree of Master of Engineering in Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY

DHAKA, BANGLADESH

MARCH 2023

DEVELOPMENT OF AN APPLICATION SOFTWARE FOR SALES PREDICTION USING MACHINE LEARNING ALGORITHMS

M. Engineering Thesis

By

FERJANA AHMED (SN. 1017140016)

Approved as to style and content by the Board of Examination on March 2023:

Dr. Md. Mahbubur Rahman Professor, Computer Science and Engineering MIST, Dhaka	Chairman (Supervisor) Board of Examination
Brig Gen Md. Mahfuzul Karim Majumder Head of the Department, Computer Science and Engineering MIST, Dhaka.	Member (Internal) Board of Examination
Col Ashfaquer Rahat Siddique, BGBMS Senior Instructor, Computer Science and Engineering MIST, Dhaka	Member (Internal) Board of Examination
Dr Kazi Abu Taher Professor, Computer Science and Engineering BUP, Dhaka	Member (External)

Department of Computer Science and Engineering, MIST, Dhaka

DEVELOPMENT OF AN APPLICATION SOFTWARE FOR SALES PREDICTION USING MACHINE LEARNING ALGORITHMS

DECLARATION

This written project submission represents the work based on proposed ideas and adequately cites and references the original source. All the principles of intellectual property, academic honesty, and integrity were maintained. There is no misinterpretation, fabrication, or falsification of any idea, data, fact, source, original work, or matter in this submission. It is recognized that any violation of the above will be cause for disciplinary action by the university and may invoke penal action from the sources which have not been properly cited or from whom proper permission has not been sought.

_____

Ferjana Ahmed

Department of Computer Science and Engineering, MIST, Dhaka.

ABSTRACT

DEVELOPMENT OF AN APPLICATION SOFTWARE FOR SALES PREDICTION USING MACHINE LEARNING ALGORITHMS

Machine learning (ML) and the use of data mining techniques are increasingly important in real-world situations. Every industry, including education, healthcare, engineering, sales, entertainment, and transportation, is benefiting from these applications' innovative nature. Due to the exponential increase of the enormous volumes of data used in commercial transactions, the business industry has significant obstacles in identifying an accurate technique and efficient prediction strategy. The conventional strategy for achieving sales and marketing objectives doesn't help businesses keep up with the pace of the competitive market since it lacks knowledge about customers' buying habits. As a result of the advancement in machine learning, significant changes are observed in the field of sales and marketing. The majority of commercial businesses rely largely on demand forecasting and knowledge of market trends. In order to improve prediction accuracy, data mining techniques are serving as efficient tools for uncovering hidden knowledge from a sizable dataset. The aim of this project is to develop a software prototype as a web service for predicting the outlet items sales of companies. The methodology of data mining with machine learning models like Linear Regression, Decision Tree, Random Forest, and XGBoost Regressor is used in this project to predict sales, and the best model for prediction is recommended based on the results analysis. Apart from the prediction, this prototype will show the graphical representation of the impact and correlations of variables as well as the outcome of the models with the predicted results. This project work will assist companies in gaining a general understanding of how to position products and outlets to give a positive customer experience that will boost sales and revenue.

সারসংক্ষেপ

DEVELOPMENT OF AN APPLICATION SOFTWARE FOR SALES PREDICTION USING MACHINE LEARNING ALGORITHMS

বাস্তব বিশ্বের পরিস্থিতিতে মেশিন লার্নিং (machine learning) এবং ডেটা মাইনিং (data mining) কৌশলগুলির ব্যবহার ক্রমবর্ধমানভাবে গুরুত্বপূর্ণ হয়ে পড়েছে। শিক্ষা, স্বাস্থ্যসেবা, প্রকৌশল, বিক্রয় সহ প্রতিটি শিল্প, বিনোদন, এবং পরিবহন খাত এই অ্যাপ্লিকেশনগুলির উদ্ভাবনী প্রকৃতি থেকে উপকৃত হচ্ছে। বাণিজ্যিক লেনদেনে ব্যবহৃত বিপুল পরিমাণ ডেটার সূচকীয় বৃদ্ধির কারণে, ব্যবসা/ শিল্পপ্রতিষ্ঠানে একটি সঠিক কৌশল সনাক্তকরণ এবং দক্ষ ভবিষ্যদ্বাণী (prediction) কৌশল প্রণয়নের ক্ষেত্রে উল্লেখযোগ্য বাধা রয়েছে। বিক্রয় (sales) এবং বিপণন অর্জনের জন্য প্রচলিত কৌশলগুলো ব্যবসায়িকদের প্রতিযোগিতামূলক বাজারের গতির সাথে তাল মিলিয়ে চলতে সাহায্য করছে না, যেহেতু গ্রাহকদের ক্রয়ের প্যাটার্ন (pattern) বা অভ্যাস সম্পর্কে এক্ষেত্রে জ্ঞানের অভাব রয়েছে। মেশিন লার্নিং এর অগ্রগতির ফলে, বিক্রয় এবং বিপণনের ক্ষেত্রে উল্লেখযোগ্য পরিবর্তন পরিলক্ষিত হচ্ছে। বেশিরভাগ বাণিজ্যিক ব্যবসার চাহিদার পূর্বাভাস (forecasting) এবং জ্ঞানের উপর নির্ভর করে বাজার প্রবণতা। ভবিষ্যদ্বাণীর (prediction) নির্ভুলতা উন্নত করার জন্য, ডেটা মাইনিং (data mining) কৌশলগুলি দক্ষ সরঞ্জাম হিসাবে কাজ করছে একটি বিশাল ডেটাসেট (Dataset) থেকে লুকানো জ্ঞান উন্মোচন করার জন্য। এই প্রজেক্টের মূল উদ্দেশ্য হল, কোন কোম্পানির আউটলেট আইটেম এর বিক্রয়ের পূর্বাভাস দেওয়ার জন্য একটি অ্যাপ্লিকেশন (application) সফটওয়্যার প্রোটোটাইপ (software prototype) তৈরি করা যা একটি ওয়েব সার্ভিস (web service) হিসাবেও কাজ করবে। এই প্রকল্পে লিনিয়ার রিগ্রেশন (linear regression), ডিসিশন ট্রি (decision tree), র‍্যান্ডম ফরেস্ট (random forest), এবং এক্সজিবুস্ট রিগ্রেসর (XGBoost regressor) এর মতো মেশিন লার্নিং মডেলগুলির সাথে ডেটা মাইনিং এর পদ্ধতি বিক্রয়ের পূর্বাভাস দিতে ব্যবহার করা হয়েছে এবং ফলাফল বিশ্লেষণের উপর ভিত্তি করে ভবিষ্যদ্বাণীর জন্য সেরা মডেলের সুপারিশ করা হয়েছে। ভবিষ্যদ্বাণী ছাড়াও, এই অ্যাপ্লিকেশন (application) সফটওয়্যার প্রোটোটাইপ (software prototype) এর মাধ্যমে ভেরিয়েবলের (variable) প্রভাব এবং পারস্পরিক সম্পর্ক এবং সেইসাথে ভবিষ্যদ্বাণী করা ফলাফলের সাথে মডেলগুলির ফলাফলের গ্রাফিক্যাল উপস্থাপনা (graphical representation) দেখানো হয়েছে। এই প্রকল্পের কাজ কোম্পানিগুলিকে একটি ইতিবাচক গ্রাহক অভিজ্ঞতা দেওয়ার জন্য পণ্য এবং আউটলেটগুলির অবস্থান সম্পর্কে একটি সাধারণ ধারণা অর্জনে সহায়তা করবে যা কোম্পানির বিক্রয় (sales) এবং রাজস্ব (revenue) বৃদ্ধি করবে।

ACKNOWLEDGEMENTS

First and foremost, I am thankful to Almighty Allah for His blessings for the successful completion of the project. My heartiest gratitude and deep respect will go to my supervisor Dr. Md. Mahbubur Rahman, Professor, Department of Computer Science and Engineering, Military Institute of Science Technology, Dhaka, Bangladesh for his supervision, affectionate guidance and proper directions whenever I sought one. It has been a privilege to work under his supervision.

Specially I am grateful to the Department of Computer Science and Engineering (CSE) of Military Institute of Science and Technology (MIST) for providing their all-out support during this project work. I am also grateful to all the members of my project committee for their valuable opinions. Their patience detecting my flaws in thinking and writing have made this project a success.

Finally, I would like to thank my family and friends for their appreciable assistance, patience, and suggestions during the course of my project, as with heavy working pressure at the office, it would be impossible for me to complete this task.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

AI	: Artificial Intelligence
BDA	: Big Data Analytics
CART	: Classification and Regression Tree Algorithm
CRISP-DM	: CRoss-Industry Standard Process for Data Mining
CV	: Cross Validation
DT	: Decision Tree
GBDT	: Gradient-boosted decision tree
LR	: Linear Regression
MAE	: Mean Absolute Error
ML	: Machine Learning
MSE	: Root Mean Squared Error
RF	: Random Forest
RMSE	: Root Mean Squared Error
SCM	: Supply Chain Management
SDLC	: Software Development Life Cycle
XGBOOST	: eXtreme Gradient Boost

CHAPTER 1
INTRODUCTION

1.1 Introduction

Business analytics are now a crucial component of every business support system because of the breakthroughs in machine learning and data analytics. In this regard, projecting sales and demand is crucial for creating business analytics solutions (Chandel et al., 2019). The basic ideas behind sellers and buyers are supply and demand. Due to the rapid expansion of malls and online shopping, competition between various shopping malls and industries is becoming more intense and fiercer daily (Behera and Nain, 2019). Every mall or store tries to give unique, limited-time deals to draw in more consumers based on the day so that the volume of sales for each item can be forecasted for inventory management of the company, logistics, and transport service, etc. In addition, for the purpose of forecasting future client demand and updating inventory management, shopping centers and various industry now keep track of the sales data of each and every individual item. These data stores essentially consist of a huge amount of client data and specific item attributes in a data warehouse. Manual infestation of this data processing 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. In spite of the significance of sales and demand forecasting, the absence of a reliable and effective demand forecasting solution results in incorrect projections that may have impact on an organization's sales and operational processes. When sales volume is overestimated, actual sales are lower than anticipated, whereas underestimated sales volume may result in higher promotional costs and a lesser profit (Bajaj et al., 2020).

For organizations to employ in their sales and operational activities, a precise sales forecast is crucial. In actuality, merchants and manufacturers may both make more successful plans for their marketing, sales, production, and procurement activities if they have a clear prediction of a product's prospective sales (Cheriyan et al., 2018). For manufacturers, wholesalers, and retailers in general, sales forecasting is an important factor to take into account. It is also a key task for many businesses engaged in supply chain activities. Sales forecasting is a complex task that becomes more challenging when there is a shortage of data, missing data values, or outliers. The most sophisticated machine learning algorithms available today provide methods for forecasting or evaluating the potential demand and revenues for a corporation. For the purpose of forecasting or predicting sales volume, many machine learning techniques are employed.

The design and improvement of market business strategies, which also aid in raising consumer awareness, are facilitated by a better prediction (Sengar and Ahmed, 2019). Predictions assist businesses in making sense of the past, spotting budgetary problems, and organizing everything. The likelihood of success rises with the creation of the strategy (N et al., 2020).

Supervised machine learning techniques can be used here to use machine learning algorithms to uncover complex patterns in the sales dynamics that include numerous risk elements as well. In this project, which focuses on sales prediction, the scenario of a one-stop-shopping mall has been explored in order to forecast the sales of various sorts of goods and comprehend the impact of various circumstances on those sales. Results with high degrees of accuracy are obtained using various components of a dataset gathered for supermarkets and the process used to create a prediction model, and these observations can be used to make decisions to increase sales.

In this project, the process will start by gathering information about a particular group of outlets operated by one of the most well-known food and beverage companies. Then, to forecast sales, XGBoost Regressor, Decision Tree, Random Forest, and Linear Regression will be employed. In the end, a collection of actual data obtained from the Companies which acts here as test dataset is used to evaluate each approach. According to the testing findings, the XGBoost Regressor provides reasonably accurate sales results. In order to better satisfy customers, enterprises will also be able to estimate item-level sales using the developed prototype software.

1.2 Scope

The scope of this project is limited to the item sales data of Food and Beverages Company. The method uses the buying behavior of customers and the collective data derived about the outlet sale of items according to the sales datasheet. Furthermore, usage of data mining techniques for data analysis and providing a predictive model as a application software using the best fit machine learning model are also covered within this project scope.

1.3 Objectives of this Project

The aim of this research is to offer a predictive model for forecasting outlet item sales. The main objectives of this project are listed below:

(a) To develop an application software for sales prediction based on user inputs.

(b) To implement machine learning algorithms on Foods and Beverages company's sales data by using the developed software.

(c) To analyze the performance of implemented models to find out maximum accuracy.

1.4 Roadmap

The rest of the project book is organized as follows.

Chapter 2: In this chapter, literature review and background information have been presented in details covering the areas of ML and sales prediction with the usage of ML. At the end of this chapter, the previously studied different works have also been highlighted.

Chapter 3: In this chapter, the methodology and architectural design of proposed system as well as implementation platform are also explained.

Chapter 4: In this chapter, the implementation of the system and processing data with model building will be focused.

Chapter 5: In this chapter, in continuation of our previous chapter, we emphasis the outcomes and analysis the results of implemented algorithm and development of software prototype.

Chapter 6: The project will be concluded in this chapter by summarizing contributions and identifying limitations. The possible future works have also discussed here.

CHAPTER 2
THEORETICAL BACKGROUND

2.1 Preliminaries of Machine Learning

This project primarily focuses on the usage of ML in the area of sales forecasting or prediction using the sales data across numerous locations and goods from the Foods and Beverages industry based on historical records. In respect to that this chapter presents the background materials along with the related works that would be required to understand the subsequent chapters where the main contributions are presented. In particular, the main target of this chapter is to discuss some important relevant topics.

The massive volume of unprocessed data being generated nowadays needs to be carefully examined in order to produce findings that are both highly informative and delicately pure in their respective industries. A branch of artificial intelligence called "machine learning" is concerned with computer programs that naturally enhance their performance over time. Machine learning is the "area of research that offers computers the ability to learn without being explicitly taught," according to Arthur Samuel, a pioneer of artificial intelligence, who first described it in 1959. To put it another way, the field of machine learning is concerned with creating algorithms that generate predictions based on data. An artificial intelligence challenge is to discover (learn) a function f: X/Y that translates the input domain X (of data) to the output domain Y (of possible predictions). The data may be trained to recognize patterns and be evaluated to make decisions with the least amount of human input using machine learning techniques. Three learning paradigms can be used to categorize machine learning: reinforcement learning, unsupervised learning, and supervised learning.

In **Supervised Learning**, a model is trained using algorithms to discover patterns in a dataset of features and labels, and the model is then used to predict the labels on the features of a new dataset. **Unsupervised learning** is the process of training an algorithm utilizing data that has neither been classed nor labeled and allowing the algorithm to act on the data without supervision. Without any prior data training, the machine's objective in this case is to categorize unsorted data according to similarities, patterns, and differences. Clustering and association are the two groups of algorithms that make up unsupervised learning. The goal of **Reinforcement Learning** is to determine how intelligent agents should behave in a given environment to

maximize the concept of cumulative reward. It is used by a variety of programs and machines to determine the optimal course of action to pursue in a given circumstance. In contrast to supervised learning, where the training data includes an answer key, reinforcement learning relies on the reinforcement agent to select how to carry out the job at hand. In supervised learning, the answer is already known, therefore the model is trained with that answer. It is obligated to gain knowledge from its experience in the absence of a training dataset. The machine learning paradigms has been illustrated in figure 2.1.

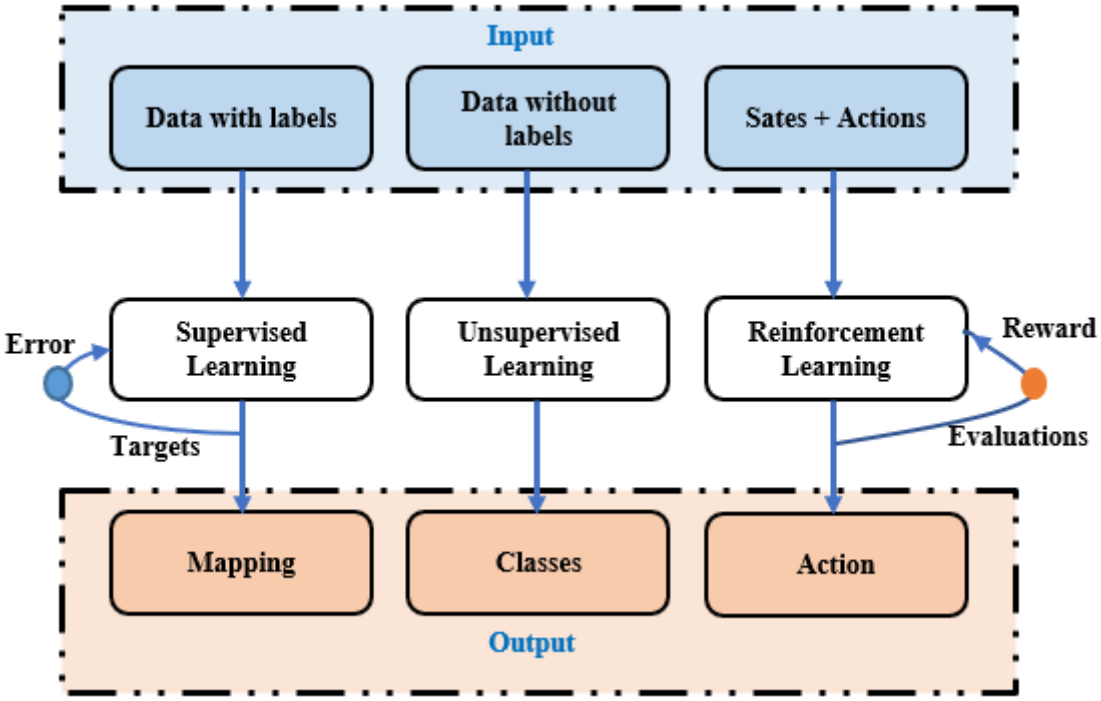


Fig 2.1: Machine learning paradigms

Due to the fact that data is currently very valuable, analysis and interpretation of data to provide effective results will also advance with technology. In machine learning, supervised and unsupervised types of tasks are both dealt with, and often a classification-type problem serves as a source for knowledge discovery. The main focus is on developing a system self-efficient so that it can perform computations and analysis to produce much more accurate and precise findings (Chandel et al., 2019). It creates resources and uses regression to make precise predictions about the future. Data can be transformed into knowledge by applying statistical and probabilistic algorithms. Sampling distributions are used as a conceptual framework for statistical.

2.2 Preliminaries of Sales Prediction

Sales prediction is the practice of projecting the volume of goods or services a sales unit will sell in the future in order to estimate future revenue. Sales forecasting is crucial in many industries and aids in increasing a company's sales by allowing for the creation of future plans. Due to the exponential increase of the vast volume of data used in e-commerce transactions, the industry has significant hurdles in identifying an accurate data mining technique and successful prediction strategy (Seyedan and Mafakheri, 2020). The ability to predict sales is a crucial requirement for effective enterprise planning and decision-making, which enables businesses to effectively plan their operations. The planning of warehouse locations by the sales and marketing department of the warehousing department is significantly influenced by sales forecasting. Equivalently, sales data can more accurately predict future sales trends. A sales forecast is, at its most basic level, a prediction of how the market will react to a company's marketing initiatives. The goal of sales prediction is to estimate future sales for businesses including supermarkets, grocers, eateries, bakeries, and patisseries.

The importance of sales forecasting extends throughout a business. Sales forecasting assists the business in reducing the stock of items whose sales are expected to decline and raising it for the goods whose sales are anticipated to rise, which will result in a rise in the firm's sales and the representation of the sales output variable. Predictions are used by finance to create budgets for recruiting new employees and planning capacity, and production schedules operating cycles using sales forecasts. The trend in sales volume over time is displayed in a company's sales analysis report. The results of a sales analysis report indicate whether sales are rising or falling. The company may evaluate the report's trends whenever it wants to choose the most appropriate course of action. Reports on sales analysis are used by the business to pinpoint market opportunities and potential growth areas. Sales analysis reports from larger organizations might only provide information for a certain region, division, or subsidiary. Actual sales are contrasted with anticipated sales in the sales analysis report. The key objective of sales forecasting is to assist the company in determining its goals and modifying its approach to increase productivity in the upcoming years.

2.3 Machine Learning Algorithms

ML is capable of handling complicated data, including time series, categorical variables, text, pictures, fuzzy elements, and other variables, from several perspectives (Cadavid, Lamouri and Grabot, 2018). However, Regression is the preferred way for solving problems like sales forecasting, and using machine learning regression algorithms can produce results that are superior to those of traditional time series methods. In comparison to more conventional time series analytical techniques, machine learning algorithms like Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost can assist uncover superior results.

2.3.1 Linear Regression

It is a supervised learning-based machine learning algorithm. It executes a regression technique. Regression uses independent variables to model a predetermined prediction value. It is mostly used to determine how variables and forecasting relate to one another. In other words, it can be referred to as a parametric technique that utilizes a set of independent variables to predict a continuous or dependent variable. This method is referred to as parametric since many assumptions are based on the data set.

It carries out the duty of predicting a dependent variable's or target variable's value (y) based on an independent variable that has been provided (x). Therefore, x (the input) and y (the output) are found to be linearly related by this regression technique (output). The straight-line equation as shown in figure 2.2 is the simplest representation of a simple linear regression equation as shown below with one dependent and one independent variable:

$$y = m * x + c$$

in which y is a dependent variable.
X is an independent variable (sale of a particular product)
M is the average of all sales (slope of the line)
C stands for the line's coefficient.

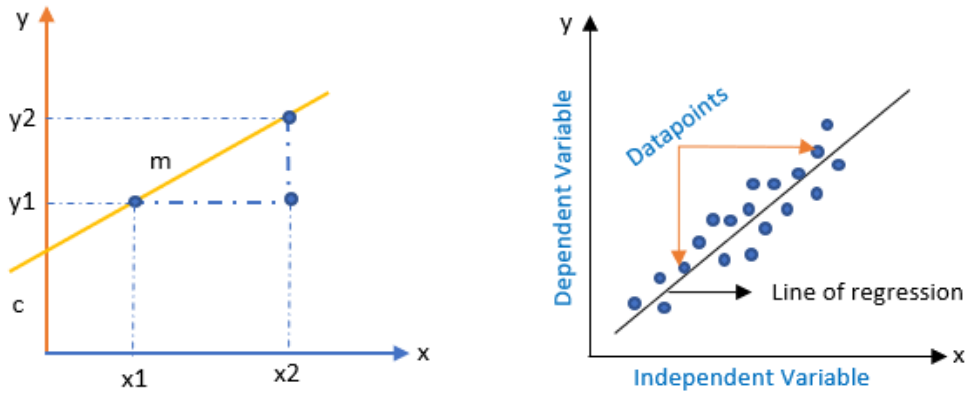


Fig 2.2: Linear Regression

Similarly, for multiple variable, the equation is represented as below:

$$y = m_1 * x_1 + m_2 * x_2 + m_3 * x_3 + \ldots\ldots\ldots\ldots + m_n * x_n + C$$

Finding the line that best fits the dependent or target variable and the independent variables of the data is the major objective of this approach. It is accomplished by identifying all m of the most ideal values. Best fit refers to a predicted value that is the closest possible to the actual data and has the least amount of error (Bajaj et al., 2020).

2.3.2 Decision tree

The decision tree algorithm is a member of the supervised learning algorithm family. Although it can be applied to problems involving classification and regression, classification problems are the ones that it is most often employed for. It can be used to form a classification tree with the root node being the first to be taken into account in a top-down fashion because it is an intuitive model with little bias. It is a well-known machine learning model. The internal nodes of the tree-structure classifier reflect the dataset features, the branches the decision rules, and the leaf nodes the results. By learning straightforward decision rules derived from previous data, a Decision Tree is used to build a training model that may be used to predict the class or value of the target variable (training data).

On the basis of the dataset's features, decision-making and testing are done in decision trees. Using the Classification and Regression Tree Algorithm (CART), a tree is constructed which is presented in figure 2.3. The root node of the tree is where the prediction process for a particular dataset's class begins. It contrasts the root attribute's values with the record/dataset attribute's values. Further division of the tree into subtrees was done on the basis of comparison

or the responses (YES/NO). As a result, it moves on to the subsequent node by following the branch that corresponds to that value. The algorithm verifies the attribute value with the other sub-nodes once again for the following node before continuing. It keeps doing this until it reaches the tree's leaf node.

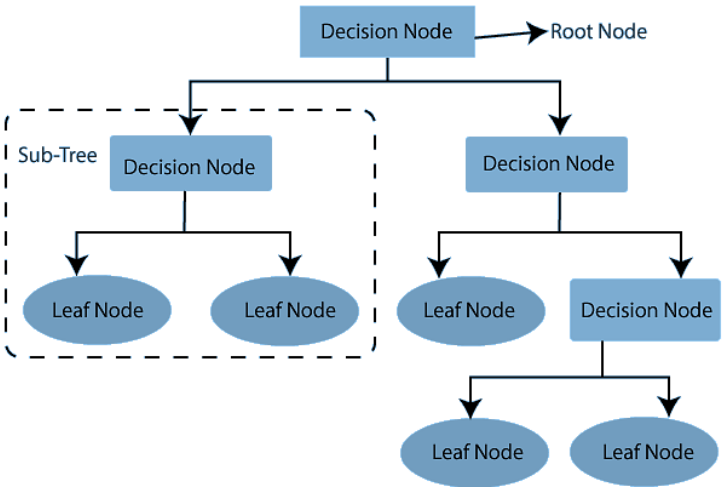


Fig 2.3: Structure of Decision Tree

2.3.3 Random Forest

Random Forest (RF) Regression, a popular statistical learning technique, takes many samples from the initial sample forecasting and combines them with decision trees to perform them. The RF algorithm uses the mean of the test predictions and can be used to perform classification, regression, and other machine learning tasks for the ensemble. This operates by building a large number of decision trees shown in figure 2.4 during the training phase, and then generating the class, which is the mean prediction (regression) or class mode (classification) of the individual trees. The issue of overfitting in decision trees can be solved by using random forests.

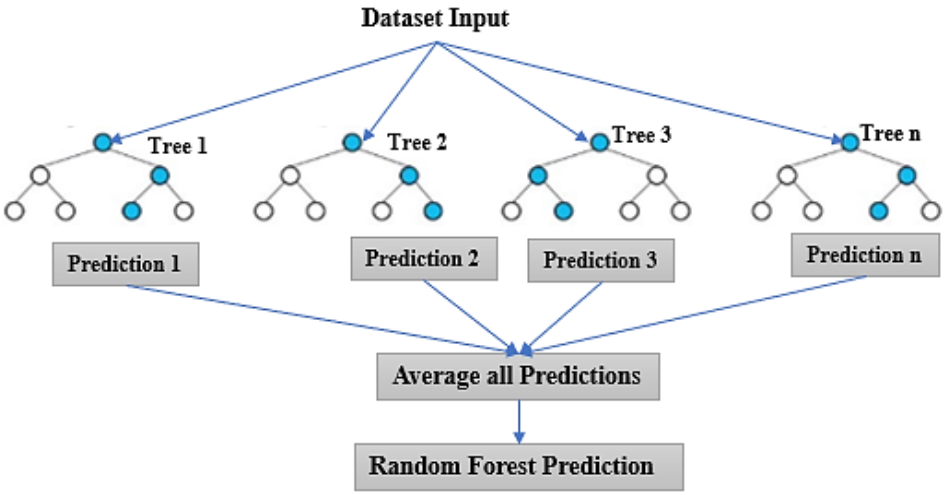


Fig 2.4: Random Forest Regression

It is feasible to describe Random Forest as a particular type of additive model that combines judgments from a series of base models to produce predictions.

To provide a forecast that is more dependable and steadier, random forest constructs numerous decision trees and combines them. While increasing the number of trees, Random Forest adds more randomization to the layout. The hyperparameters of Random Forest for bagging are similar to those of a decision tree or a classifier. When breaking a node, it searches for the best feature from a random subset of characteristics rather than the most suitable feature. In general, a better model is produced as a result of the great diversity this produces. Because it only chooses a small portion of the features from each node to break, random forest is popular for its faster prediction performance and reduced memory usage.

2.3.4 XGBoost Regressor

The distributed gradient boosting library, XGBoost, is an optimized and highly accurate implementation constructed primarily to enhance the performance and computational speed of machine learning models. It is intended to push the limits of processing power for boosted tree methods. Under the Gradient Boosting framework, it implements machine learning techniques. In the same manner as GBDT/GBM, XGBoost offers parallel tree boosting. Utilizing a level-wise approach, it assesses the quality of splits at each potential split in the training set by scanning through gradient values and using these partial sums.

Gradient boosting is a variation on the concept of "boosting," in which new models are built to forecast the errors or residuals of older models, which are then combined together to provide

the final prediction. Since a gradient descent approach is used to reduce loss when introducing new models, it is an extension of boosting. This strategy handles challenges involving predictive modeling for both regression and classification. It's possible that this is an ensemble learning technique that integrates numerous decision trees to create a final prediction model. This approach is based on the idea that a group of weak learners can, through the boosting process, be united to create a strong learner.

A group of deep decision trees are trained iteratively by GBDTs, which use the error residuals from one model's previous iteration to fit the model that comes after it. The weighted average of all the tree predictions represents the final prediction.

Let h(x) be the new tree that was added to the model and let F(x) be the complete model after the t-1 round as represented in below equation:

$$F0= 0$$

$$Ft\ (x) = Ft\text{-}1(x) + h(x)$$

Each new function is an effort to fix the errors in the model created in earlier iterations. Therefore, the residual Ft-1 (x) must be predicted by the new function(x) as per figure 2.5.

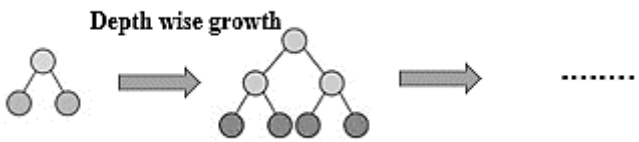


Fig 2.5: XGBoost basic architecture

2.4 Related Works

In this project, variety of data mining approaches has been used to forecast retail sales. The relevant work is also discussed in this section to help us become comfortable with the task, which entailed estimating the sales on any given day at any retailer. Many writers have used machine learning algorithms to predict sales as well as analyze predicted sales. The sales forecasting research is briefly summarized in this section.

Using historical sales data, Cheriyan et al. (2018) examined the benefits of ML algorithms over conventional data mining techniques. The findings are summed up in terms of the dependability and accuracy of effective prediction and forecasting systems. According to the investigations, the Gradient Boost algorithm provides the best fit and displays the highest level of classification matrix accuracy.

The predicting problems in online business transactions were examined by Chandel et al. (2019), who also suggested a stacked generalization method made up of sub-level regressors. After that, it put the general model and each classifier's results to the test. When additional data is used, this method will make significantly better predictions. The variation between the suggested model and random forest is not statistically significant, hence the proposed method can forecast demand because it is accurate with fewer data. The sales of the big retail companies are forecasted using the XGBoost regressor and gradient-based algorithm.

Seyedan and Mafakheri (2020) give a review paper that lays the way for a critical debate of BDA applications in SCM by emphasizing a number of important discoveries and outlining the current difficulties and gaps in BDA applications for demand forecasting in SCs. A variety of directions for more research are then presented in the paper's conclusion in light of this.

Bajaj et al. (2020) proposed another machine learning algorithm to envisage the pattern of sales and the quantities of the products to be sold keeping in view the sales of previous years. Random Forest model was found as a suitable method for their study.

In order to pre-process raw data collected from a big mart for missing data, anomalies, and outliers, Meghana et al. (2020) used the random forest and XG booster methods. The findings were then predicted using an algorithm, and the best model was predicted after a comparison of all the models. The random forest approach and XG Booster technique are the best models for forecasting sales in the large mart, according to their analysis of the accuracy of the predictions made by the various models.

Sengar et al. (2019) analyze the optimal form in order to forecast the stock market's rate. Diverse presentations and objects must be considered as details in the investigation process, and processes cannot be entirely divided. Therefore, all of the data was reprocessed and then modified for analysis. Support vector machine and random forest are the two algorithms that were used in this paper. By comparing the accuracy of the two methods, it was discovered that the random forest algorithm performed best for market price prediction.

Sadia et al. (2019) analysis of the best form predicts the stock market's pace. Support vector machine and random forest are the two algorithms that were used in the investigation procedure. By comparing the accuracy of the two algorithms, it was discovered that the random forest method performed best for market price prediction.

The need for the organization to invest in forecasting approaches over conventional methods is covered in the paper published by Behera and Nain (2019) along with a study of machine learning trends utilized in the demand and sales forecasting sector from 2009 to 2017.

Punam et al. (2018) introduce the prediction of sales of goods by two-level approaches, which is a statistical model that makes use of MAE values. In that case, a variety of algorithms, including KNN, Support Vector Regression, Linear Regression, and Regression Tree, are used to evaluate the Squared Error Value.

A software solution for predicting future sales based on historical sales data is suggested by Kadam et al. (2018) Prior to training, the raw data is first evaluated and preprocessed. To get to a conclusion, the output of two algorithms is compared. These include multiple linear regression and random forest. With the help of random forest and multiple linear regression models, this approach is utilized to forecast future sales for large retail companies. Iakovou, Kanavos and Tsakalidis (2016) have presented a study on modeling and forecasting consumer behavior using data on supermarket merchandise. They explicitly suggested a new approach for product recommendations based on an analysis of each customer's purchases.

Cadavid, Lamouri and Grabot (2018) attempted to examine representative ML applications in SCM, with a focus on the particular field of Demand and Sales Forecasting (D&SF) using conventional techniques. In the field of finance, Shah, Isah and Zulkernine (2019) offered a succinct overview of stock markets and a taxonomy of stock market prediction methodologies in stock analysis and prediction.

The authors (Dairu and Shilong, 2021) target forecasting the future costs of valuable metals like diamond, utilizing Gradient Boost algorithm, intending to get the most precise consequence of all. XGBoost includes an approximation method for exact greedy algorithms, in-memory data storage for parallel learning, cache-aware prefetching, and out-of-core processing. With XGBoost, users may manage a bigger dataset and run it more quickly.

CHAPTER 3
METHODOLOGY AND DESIGN

3.1 Methodology Overview

This section will cover the methodology and proposed architectural design used to build the predictive models required for Sales Forecasting.

Data science projects, like predictive models, do not have a lifecycle with well-defined steps like Software Development Life Cycle (SDLC). Usually, data science project workflow contains the repeated hold-ups and iteration. It's a very time-consuming process. Even if in a real-life business scenario, it takes months, even years, to get to the endpoint where the developed model starts to show results. As part of the implementation of this project, a standard workflow of a data science project which is based on one of the oldest and most popular-CRISP DM is followed. CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology provides a structured approach to planning a data mining project which is now broadly adopted in the genre of data science and ML. Using this methodology, the architectural diagram of the proposed system is designed to produce an effective data product whereas this data product is a dashboard to facilitate decision-making to solve a business problem. However, to reach the end goal of producing data products, step by step workflow process is illustrated in figure 3.1:

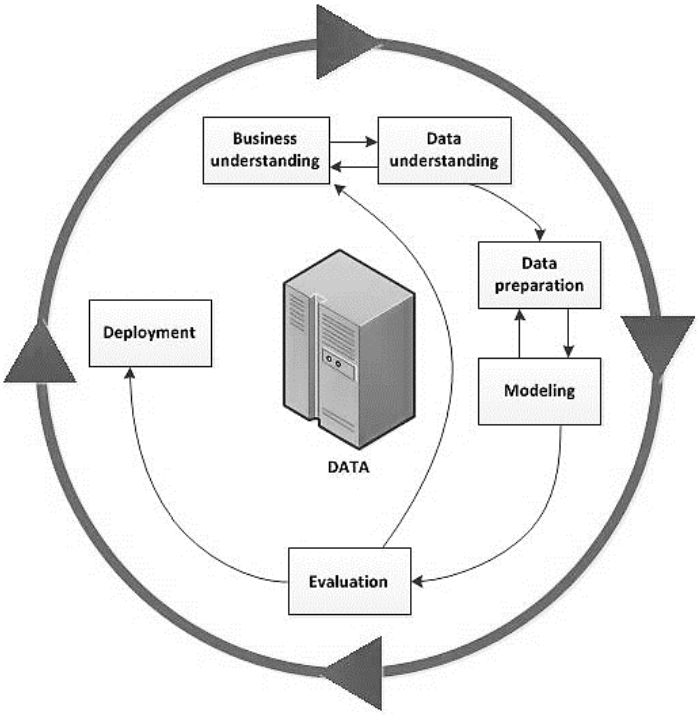


Fig 3.1: Data Mining process model using CRISP-DM methodology

3.2 Proposed System Architecture

Following the described methodology of CRISP-DM, the proposed work suggested the following architecture shown in figure 3.2 to forecast sales of different items at different outlet locations. Figure depicts the suggested system's architecture diagram.

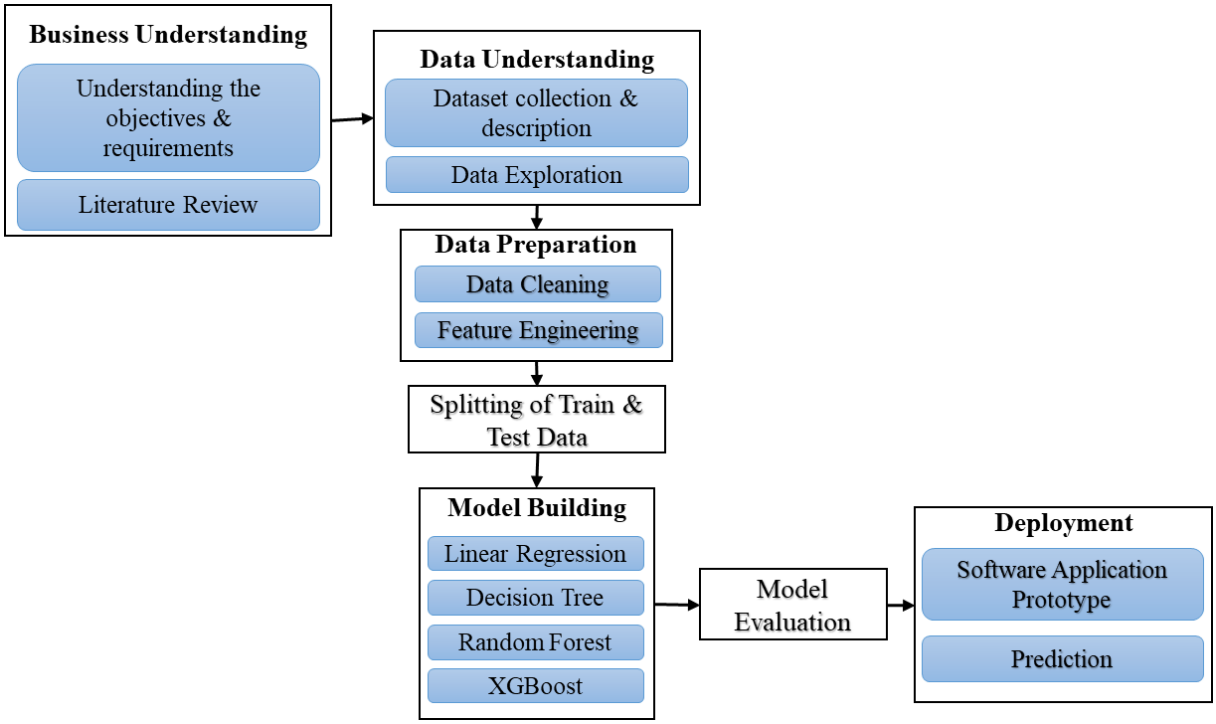


Fig 3.2: Architecture of proposed system

As part of the proposed architecture, the project begins by understanding the business context and identifying data requirements. Relevant data from the Kaggle repository would be collected with the train and test datasets. After that, these datasets go through a preliminary data exploration analysis which includes univariate and bivariate analysis. In later stage data preparation consisting the method of pre-processing is performed which takes care of missing, error values and outlier correction in the dataset. selected features are also modified in this stage. furthermore, feature transformation is also conducted for creating new features from existing features. Lastly, dataset preparation is completed for the deployment of ML Algorithms. The selected four algorithms would then be applied and trained using the train data to develop the predictive models. This performance of the models would be tested using the test data to determine the model with the most accurate predictive performance. At last, the deployment part, software prototype will be created to integrate with the model. Best fit model

will be chosen for the system deployment. In that, the data will be visualized from the interface and any user can show the item sales inserting the input of item and outlet identifier.

3.2.1 Business Understanding

The objectives and requirements of the project are the main emphasis of the business understanding phase. The areas of the business that need to be explored must be determined, a business model must be created if necessary, the analysis' expectations must be outlined, the goal must be defined, a hypothesis must be formed, and resources for data collection must be located. The objective of this project is to develop a prediction model to determine the sales of each product at a certain store. Our area of study in this project is using the sales data. As a result, we'll strive to comprehend the characteristics of the goods and the establishments that are important for boosting sales. In this step, a number of hypotheses like product level or store level are developed by looking at the project goal. As per the Initial understanding, product level hypothesis can be with the parameter of product brand, offers, advertisements, usage, packing process whereas store level hypothesis can be generated with the location of store, population in the location of store, type of store location, size of store, customer relationship with store people, advertisements, local need fulfillment, competitors etc. which will affect the sales. To simplify this hypothesis, one example can be presented for better understanding. For example, branded product can be sold higher in one store basing on the customer’s requirement as per store location. But it will not be assumed for all stores. Similarly, many more hypothesis can be generated for better understanding of the requirements.

3.2.2 Data Understanding

The Data Understanding phase builds on the Business Understanding phase's foundation by emphasizing the identification, gathering, and analysis of data sets that can be useful to the project. The necessary information is obtained either through web scraping or from reputable sources like GitHub or Kaggle. Selecting the appropriate data is crucial since it will reveal whether the initial hypothesis was correct. This phase also includes data description, which includes the characteristics of the data, such as its format, quantity of records, or field identifiers, as well as data exploration, which includes data analysis, visualization, and the identification of correlations between the data. In this step, the accuracy of the data obtained

from multiple sources is also confirmed, and the business objectives are taken into consideration in order to comprehend the data.

3.2.2.1 Exploratory Data Analysis

An exploratory analysis must be carried out in order to clearly comprehend the nature of the chosen data. A strategy for analyzing datasets using statistical graphics and other data visualization techniques is called Exploratory Data Analysis (EDA). It summarizes main characteristics or features of data, variables and corresponding relationships. The exploratory analysis includes two types of analysis i.e., univariate which deals with only one attribute and bivariate which deals with two attributes that are conducted on data, to summarize and find patterns in the data Target Variable.

A. Univariate Analysis

The simplest types of data analysis, known as univariate analysis, focus on one variable (or data column) at a time. As part of the Univariate analysis, to explore all the individual variables, count plots are being used to gain some insights with data visualization.

B. Bivariate Analysis

Any concurrent relationship between two variables or qualities is a definition of Bivariate Analysis. It investigates how two variables interact as well as any differences between the two variables and their possible causes.

Univariate and Bivariate Analysis on selected dataset are briefly discussed in chapter – 4 with the visual explanation.

3.2.3 Data Preparation

This phase involves three sub-steps: selection, pre-processing, and transformation, and it prepares the final data set(s) for modeling. To start, choose the data for analysis by examining their uniformity, reasonableness, and consistency. The data is then assessed during the pre-processing stage and, depending on its importance, is either included in or eliminated from the main data that will be studied.

The chosen sales dataset may need to go through a number of processes, each of which is essential for testing the accuracy of the sales forecast, in order to build a model that can predict results. Raw data is the foundational stage of all data and the original source of judgments based on data. Data must be developed before being used in machine learning algorithms because it cannot be used in its natural form due to the manner it was gathered. Visualization and the creation of analytical statements are not possible without processing the raw data. Prior to making any choices regarding its utilization or not, the objective is to ascertain what kind of information is required.

3.2.3.1 Data Cleaning

For the purpose of developing models, data must be properly prepared and cleansed. The acquired data is placed through a "cleaning" process to make sure it is properly categorized and that any discovered information gaps are filled with the pertinent data. In a nutshell, data cleaning involves correcting or eliminating inaccurate, damaged, improperly formatted, redundant, or incomplete data from a dataset. Any of the following 4 steps can be mapped as a component of this cleaning procedure shown in figure 3.3.

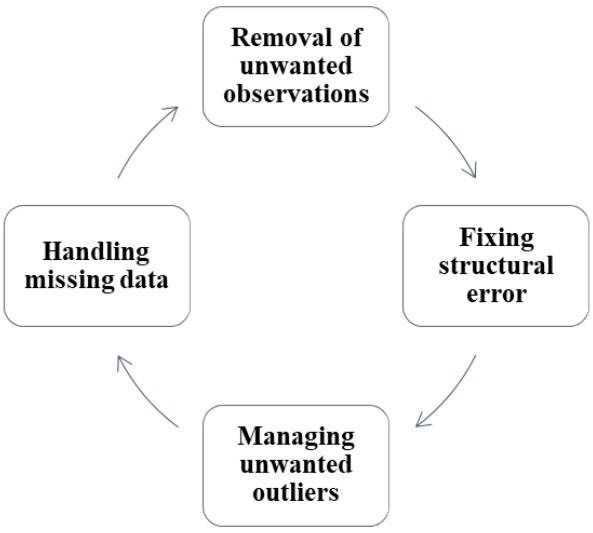


Fig 3.3: Data Cleaning process

The process of developing predictive models can be severely hampered by missing data because those values may contain crucial information that would enable more accurate forecasts. Therefore, performing missing data imputation becomes crucial. Depending on the issue and the data, many approaches can be used to handle missing values. Here are a few of the standard methods:

(a) **Row deletion:** Any observations in a dataset that have missing values for any variable are eliminated. Loss of information and a decrease in the model's predictive power are two disadvantages of the strategy.

(b) **Mean/Median/Mode Imputation:** When a variable is continuous, missing values can be imputed using the mean or median of all previously recorded values for the variable. In the case of categorical variables, the mode of the available values might be used to replace the missing values.

(c) **Creating a Prediction Model:** A predictive model can also be created to impute missing data in a variable. In this case, the missing data variable will be used as the target variable while the other variables will serve as predictors. Two datasets from our data can be generated, one without missing values for that variable and the other with missing values. The training set from the first set would then be used to apply the predictive model to the second set in order to forecast the missing value.

3.2.3.2 Feature Engineering

The feature engineering process is a key component of all machine learning models. Raw data is converted into features that can be used in machine learning algorithms, such as predictive models, during the preprocessing processes. As outcome variables and predictor variables make up predictive models, the most practical predictor variables are produced and chosen for the predictive model during the feature engineering phase. The four major processes of feature engineering in machine learning (ML) are feature creation, feature transformation, feature extraction, and feature selection. During the data exploration phase, our data set may undergo some modification. These sorts of fluctuation from the data set are addressed through feature engineering, which improves the predictive model.

3.2.3.3 Encoding Categorical Values

For the correlation and regression analysis, converting categorical variables to numerical variable is required. To increase the effectiveness of the Machine Learning model, the categorical data can be transformed into numerical data. For this procedure, the two following methods will be used:

(a) **Label encoding**: This method assigns a number to each category in a variable. Ordinal variables under categorical values are better suited to it.

(b) **One hot encoding**: This technique involves converting each category of a categorical variable into a fresh binary format (1/0) and adding it as a feature. Comparing each level of the numerical variable with a predetermined starting point is one of the most widely used techniques.

Both of the aforementioned methods will be used in this project to encode categorical values for the data set that was used.

3.2.3.4 Data Correlation

Data correlation is a technique for determining the link between variables and for making predictions about one attribute based on another attribute. The correlation will be positive if increasing one feature results in a rise in another feature, and negative if increasing one feature results in a decrease in another. There is no correlation if there is no relationship between any two attributes.

The attributes with negative correlations can be eliminated to increase the effectiveness of the machine learning model. It is a statistic that measures the linear correlation between the two variables X and Y. The range of correlation values is from -1 to 1.

The following correlations exist:

(a) negative correlation: 0 and >= -1

(b) positive correlation: > 0 and = 1

(c) no correlation: 0

The correlation plot demonstrates the relationships between each and every pair of potential variables in the data. The stronger the connection between the variables, the more important it is to take into account the correlation.

3.2 4 Splitting of Train and Test Data

Fitting data into the algorithm and teaching it to spot patterns is one of the most crucial steps. Once the model has learned the pattern, another dataset must be fed to it in order to predict the

result. Two distinct datasets are not imported for the train and test portions of this project in order to prevent overfitting. Thus, splitting is carried out within a single dataset. Furthermore, the ratio of spitted data can be readjusted also for getting the better accuracy.

3.2.5 Model Building

Machine Learning prediction models are trained in this process and then later on evaluated using the data. When the data set is ready for usage, a predictive machine learning model is created by learning from the training data. Either developing a classification model (if the goal variable is qualitative) or a regression model (if the target variable is quantitative) depending on the data type of the target variable.

3.2.6 Model Evaluation

Model evaluation helps in examining which algorithm is most appropriate for the given dataset in a specific problem. This approach evaluates the performance of various machine learning models using the same input dataset. The technique of evaluation emphasizes the importance on how well the model predicts the final outcome.

3.2.6.1 Model Evaluation Techniques

The two categories of techniques used to evaluate a model's performance are holdout and cross-validation. In the holdout approach, the dataset is partitioned into three subsets at random: the train, validation, and test datasets. While in cross-validation, the initial observation dataset is divided into a training set (used to train the model) and an independent set (used to evaluate the analysis). The most popular cross-validation method uses k-fold cross-validation, which divides the original dataset into k folds of equal size. The k is a user-specified number that is typically recommended to have a value of 5 or 10. This is performed k times, with each time using one of the k subsets as the test set/validation set and combining the remaining k-1 subsets to create the training set. To determine our model's overall efficacy, average of the error estimation across all k trials will be calculated.

3.2.6.2 Model Evaluation Metrics

Metrics for model evaluation are needed to measure model performance. Based on a specific machine learning task (such as classification, regression, ranking, clustering, topic modeling,

among others), the evaluation metrics are chosen. Classification Accuracy, Confusion Matrix, Logarithmic Loss, Area Under Curve (AUC), and F-Measure are the evaluation measures for the classification task. Whereas Mean Absolute Error and Root Mean Squared Error are the most frequently used measures for evaluating regression specific problems (Narkhede et al., 2020). The problem statement for this project is a regression task. As a result, the model will be examined using metrics for regression.

A. Mean Absolute Error (MAE)

It represents the mean of the absolute differences between the model's forecasted value and the actual value as shown below:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

where, **N** = total number of data points, y_i = actual value, $\hat{y}_i$ = predicted value

B. Root Mean Squared Error (RMSE)

The Root Mean Square Error (RMSE) is the square root of the difference between the given data's expected and actual values. The Root Mean Square Error can be calculated by taking the square root of the MSE. It is the most common metric evolution method applied to regression issues. It is predicated on the idea that errors are unbiased and has a normal distribution. When the RMSE is higher, there are significant differences between the expected and actual values. The mathematical representation has been shown below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}$$

where, **n** = total number of data points, y_j= actual value, $\hat{y}_j$= predicted value

C. R squared (R^2)

R-squared calculates the percentage of the dependent variable's variance that is accounted for by the independent variable. A common metric used to assess model accuracy is R squared. It indicates how closely the data points match the fitted line produced by a regression method. A better fit can be determined by a higher R squared value. This aids in establishing the relationship between the independent and dependent variables.

R^2 scores varies between 0 and 1. The R^2 value should be as near to 1 as possible. If R^2 is equal to 0, the model is not outperforming a random model. The regression model is incorrect if R^2 is negative. It is the ratio between the squares' sum and their combined sum as shown in below:

$$R^2 = 1 - \frac{SSE}{SST}$$

where **SSE (Sum of Squared Errors)** is the sum of the squares of the differences between the actual value and the forecasted value shown in the following equation:

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Total sum of squares, or **SST**, is the sum of the squares representing the difference between the actual value and its mean as presented in below equation:

$$SST = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

Here, y_i is the actual target value that was seen, $\hat{y}_i$ is the actual value that was predicted, and y-bar is the mean value. ***m*** is the total number of observations.

3.2.6.3 Feature Importance

The concept of "feature importance" refers to the process of giving values to the features that make up a predictive model, which assesses how beneficial a given variable is to the model and prediction being used at the time. The feature importance scores reveal information about a given model, including which features are most and least crucial to the model's ability to make predictions. It could be applied to improve a predictive model. Using the importance scores, one may choose which features to keep (those with higher ratings) and which to discard. This kind of feature selection can make the modeling problem simpler, speed up the modeling process, and in some situations, enhance model performance.

3.2.7 Deployment

If the customer cannot access the results of the model, it is not very useful. Therefore, deployment should be seen in terms of what is necessary to really utilize the project's results. The report may be straightforward or sophisticated based on the nature of project, such as implementing a real-time, live predictive model. In this phase, a deployment plan is developed,

final reports are made, and the entire process is examined to search for errors and determine whether any processes need to be repeated. This strategy is established to track and manage the outputs of the data mining model in order to assess their utility.

3.3 Implementation Platform and Language

The process of analyzing big data and extract knowledge-rich information from that data is challenging. Therefore, in order to extract the information from complicated datasets and make better decisions going forward, there is a need of strong and efficient data mining tool. R is an effective and cost-free open source data mining technology that are employing in this project. R comes with a number of built-in packages like ggplot2, VIM, corrplot, data.table, dplyr, caret, and others that help us be more productive. R is an open-source programming language and environment for data analysis. Data analysis is a crucial component of statistics and involves the process of converting data into knowledge, insight, and understanding. It enables users to carry out a variety of crucial operations for the efficient processing and analysis of big data. R has a wide range of pre-built statistical modeling techniques and machine learning tools that let users build data products and conduct reproducible research. Despite the fact that other programs can also be used to handle large amounts of data, R really stands out when it comes to data analysis because of the vast array of third-party algorithms and built-in statistical formulae it offers. Data processing, comparison of several model possibilities, and result visualization are crucial for developing a robust and trustworthy statistical model. From the literature review, For more noticeable results, Shinde et al. (2017) introduced Random Forest (RF) and Latent Dirichlet Allocation (LDA) in the R package. By analyzing fertility-related large data and building a mathematical model to more accurately anticipate these possibilities, the author conducted case studies to demonstrate the theory in a realistic manner. The interactive nature of the R language encourages investigation, explanation, and presentation, which is why it has become so widely used.

An integrated development environment (IDE) for R and Python is called RStudio. It has a console, a syntax-highlighted editor that allows for direct code execution, tools for graphing, history, debugging, and workspace management. There are desktop versions of RStudio that are both open source and commercial (Windows, Mac, and Linux). In this project, The RStudio Desktop version has been used which was Open source. In addition to that, as part of the development of application software prototype, the project's graphical user interface was

created using R Shiny. In order to provide an elegant and simple-to-use web framework for creating web apps in R, the team at RStudio, PBC created the open source Shiny package. With the help of Shiny, R users can build amazing apps, dashboards, and interactive maps. It enables the creation of interactive web page applications directly from R without the need for prior knowledge of HTML, CSS, or JavaScript. After loading the dataset into RStudio, data exploration, visualization, cleaning, imputation, and prediction are carried out. A machine with Windows 10 pro, 16GB of RAM, and 256 GB SSD is used to conduct the experiment.

CHAPTER 4
IMPLEMENTATION

4.1 Dataset Collection and Description

This chapter provides an overview of the detail analysis of dataset with pictorial representation as well as the implementation of model and solution prototype. This chapter sequentially emphasizes on the construct of the data set and the mechanism of the experiments, analysis of the data with data modeling tasks, the graphical user interface of solution prototype and performance evaluation of the predictive models developed by the algorithms.

The dataset used for this project consisted of item sales of different outlets of supermarkets. The data was derived from an online open-source data repository, Kaggle. The dataset comprises both categorical, numeric input variables, and a continuous output variable, hence the regression task. The dataset comprises train and test sets. There are 8523 different items in the train set, each with 12 attributes (11 input variables, and an output variable). The products are spread throughout different cities and locations. The test set contains 5681 observations with 11 attributes. The sales for the test dataset is to be predicted. The feature list of the dataset has been shown in Table 4.1:

Table 4.1: All attributes in the dataset

1. Item_Identifier - A special identifier for every product.
2. Item_Weight – Weight of the product.
3. Item_Fat_Content – The product's fat content.
4. Item_Visibility – Percentage of a store's overall display space devoted to a product.
5. Item_Type – Category of products.
6. Item_MRP – The item's list price.
7. Outlet_Identifier - Individual identification number for every store.
8. Outlet_Establishment_Year – The founding year of each store.
9. Outlet_Size - The store's size.
10. Outlet_Location_Type - The classification of the city where the store is situated.
11. Outlet_Type - Whether the outlet is a supermarket or a grocery shop.
12. Item_Outlet_Sales - Product sales for each store.

The attribute Item Outlet Sales, which is the target variable that will forecast out of these attributes, is the response variable, while the other attributes are utilized as the predictor variables. This dataset contains a wide range of underlying patterns shown in figure 4.1 that shed light on the subject forecast.

```
$ Item_Identifier          : chr  "FDA15" "DRC01" "FDN15" "FDX07" ...
$ Item_Weight              : num  9.3 5.92 17.5 19.2 8.93 ...
$ Item_Fat_Content         : chr  "Low Fat" "Regular" "Low Fat" "Regular" ...
$ Item_Visibility          : num  0.016 0.0193 0.0168 0 0 ...
$ Item_Type                : chr  "Dairy" "Soft Drinks" "Meat" "Fruits and Vegetables" ...
$ Item_MRP                 : num  249.8 48.3 141.6 182.1 53.9 ...
$ Outlet_Identifier        : chr  "OUT049" "OUT018" "OUT049" "OUT010" ...
$ Outlet_Establishment_Year: int  1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...
$ Outlet_Size              : chr  "Medium" "Medium" "Medium" "" ...
$ Outlet_Location_Type     : chr  "Tier 1" "Tier 3" "Tier 1" "Tier 3" ...
$ Outlet_Type              : chr  "Supermarket Type1" "Supermarket Type2" "Supermarket Type1" "Grocery Store" ...
$ Item_Outlet_Sales        : num  3735 443 2097 732 995 ...
```

Fig 4.1: Structure of data

Sample preview of train and test datasets are shown in the table 4.2 and 4.3:

Table 4.2: Preview of Train Dataset

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
FDA15	9.3	Low Fat	0.016047301	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.138
DRC01	5.92	Regular	0.019278216	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
FDN15	17.5	Low Fat	0.016760075	Meat	141.618	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.27
FDX07	19.2	Regular	0	Fruits and Vegetables	182.095	OUT010	1998		Tier 3	Grocery Store	732.38
NCD19	8.93	Low Fat	0	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052
FDP36	10.395	Regular	0	Baking Goods	51.4008	OUT018	2009	Medium	Tier 3	Supermarket Type2	556.6088
FDO10	13.65	Regular	0.012741089	Snack Foods	57.6588	OUT013	1987	High	Tier 3	Supermarket Type1	343.5528
FDP10		Low Fat	0.127469857	Snack Foods	107.7622	OUT027	1985	Medium	Tier 3	Supermarket Type3	4022.7636

Table 4.3: Preview of Test Dataset

Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	utlet_Type
FDW58	20.75	Low Fat	0.007564836	Snack Foods	107.8622	OUT049	1999	Medium	Tier 1	rmarket Type1
FDW14	8.3	reg	0.038427677	Dairy	87.3198	OUT017	2007		Tier 2	rmarket Type1
NCN55	14.6	Low Fat	0.099574908	Others	241.7538	OUT010	1998		Tier 3	rocery Store
FDQ58	7.315	Low Fat	0.015388393	Snack Foods	155.034	OUT017	2007		Tier 2	rmarket Type1
FDY38		Regular	0.118599314	Dairy	234.23	OUT027	1985	Medium	Tier 3	rmarket Type3
FDH56	9.8	Regular	0.063817206	Fruits and Vegetables	117.1492	OUT046	1997	Small	Tier 1	rmarket Type1
FDL48	19.35	Regular	0.082601537	Baking Goods	50.1034	OUT018	2009	Medium	Tier 3	rmarket Type2
FDC48		Low Fat	0.015782495	Baking Goods	81.0592	OUT027	1985	Medium	Tier 3	rmarket Type3

4.2 Data Analysis and visualization

All the graphs in this project were produced using the ggplot2 tool. Target variable Item Outlet Sales may be seen by visualizing its histogram because it is a continuous variable. From Figure 4.2, Item Outlet Sales is a right-skewed variable, so treating its skewness would require some data processing.

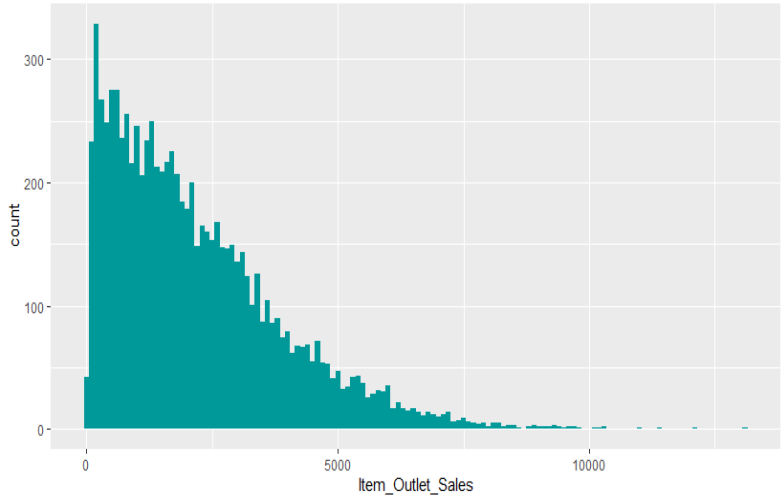


Fig 4.2: Plot of Target variable Item_Outlet_Sales

4.2.1 Univariate Analysis

The independent numeric and categorical variables from the dataset are plotted using their histograms to visualize their distributions.

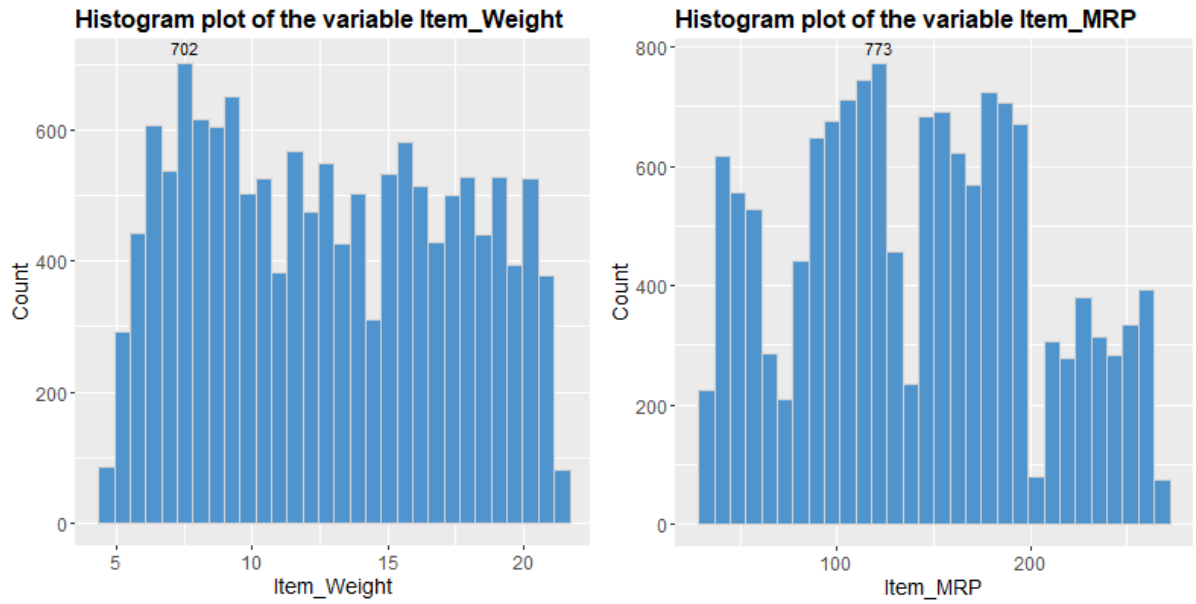
A. Distribution of Independent Numeric Variable

Following observations can be drawn from the distribution of Independent Numeric Variable shown in figure 4.3.

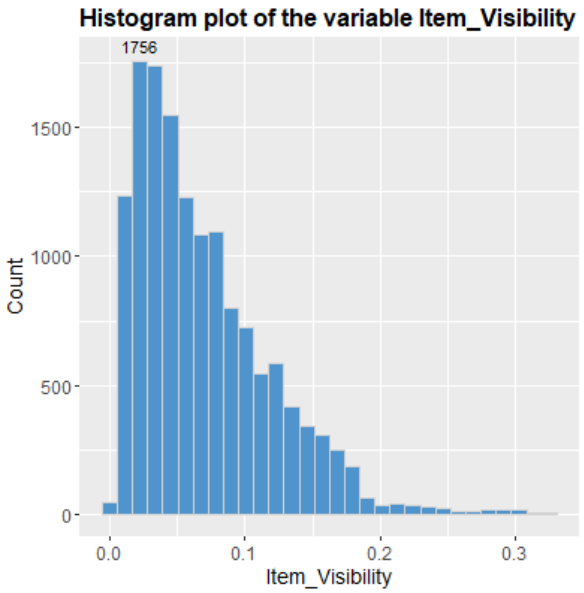
(a) Item_Weight – Item_Weight hasn't shown any obvious patterns. Item_Outlet_Sales are randomly distributed throughout the full Item Weight range.

(b) Item_MRP – Item_MRP has four different distributions.

(c) Item_Visibility - Right-skewed pattern in Item Visibility, with a minimum value of 0 having the maximum number of counts.



(a) (b)



(c)

Fig 4.3: Distribution of Independent Numeric Variable (a) Item_Weight, (b) Item_MRP, (c) Item_Visibility

B. Distribution of Independent Categorical Variable

Only a finite set of values can be assigned to an independent categorical variable or feature. Item_Fat_Content's count plot, which includes the two categories Low Fat and Regular Fat written under separate identifier names, is shown. It has been noted that majority of the items contain low fat content. Even though there are two different sorts of fat content, they are

mentioned by various ways such as ‘reg’ rather than ‘Regular’ and ‘low fat’, ‘LF’ rather than ‘Low Fat’. Similar categories of Regular and Low-Fat item have been combined and same figure plotted again as shown in figure 4.4 and 4.5 respectively. It is noted that there are more Low fat products available than regular ones.

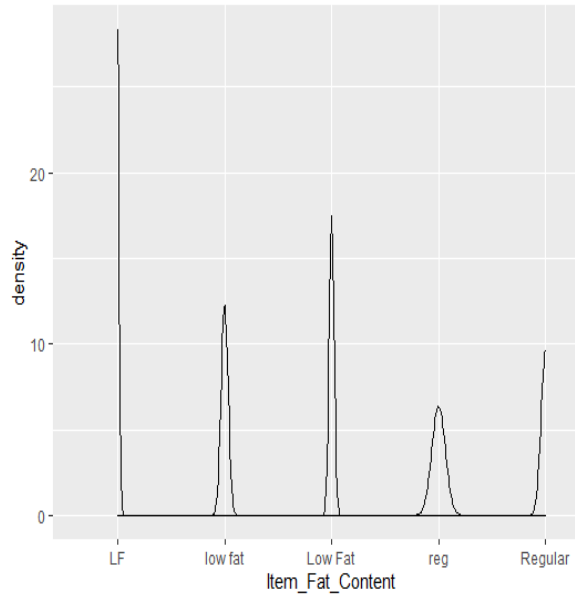


Fig 4.4: Plot for Item_Fat_Content

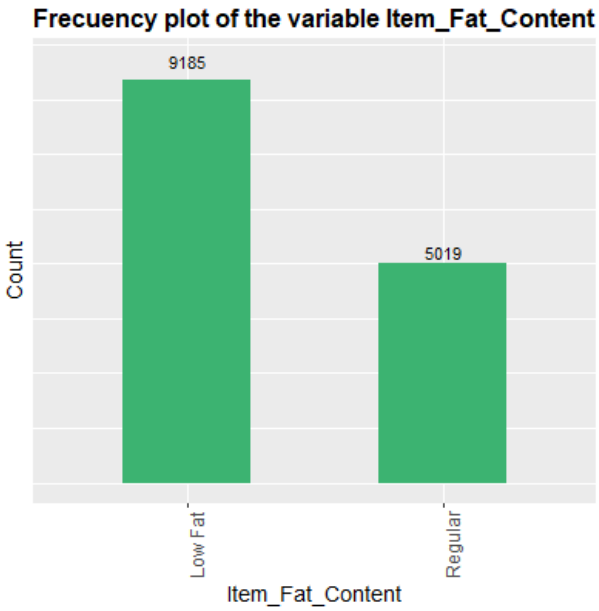


Fig 4.5: Plot for combined Item_Fat_Content

The count plot for all categorical variables shown in figure 4.6 depicts the following realizations-

(a) The plot for Outlet_Size shows that 4016 of the data in Outlet_Size are either blank or missing. The missing values in the Outlet_Size can be handled using bivariate analysis. In addition, only a small percentage of the outlets are high or huge in size.

(b) It is also evident from the count of this distribution of Item_Type that majority of the items are ‘Fuits and Vegetables’ with "Snack Foods" having the second-highest item count. On other hand, ‘Seafood’ is the least counted item.

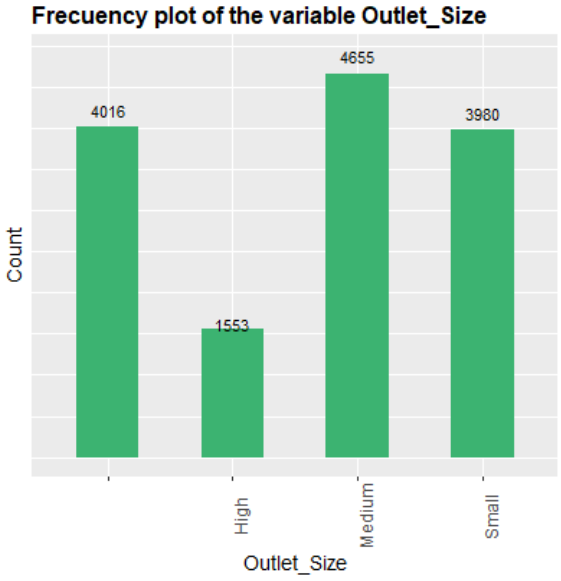
(c) Highest no of outlets is OUT027 and OUT013 sequentially and OUT019 contains the lowest no outlets which is visible from Outlet_Identifier plot.

(d) Maximum outlets are Supermarket Type1 in the classification of Outlet_Type.

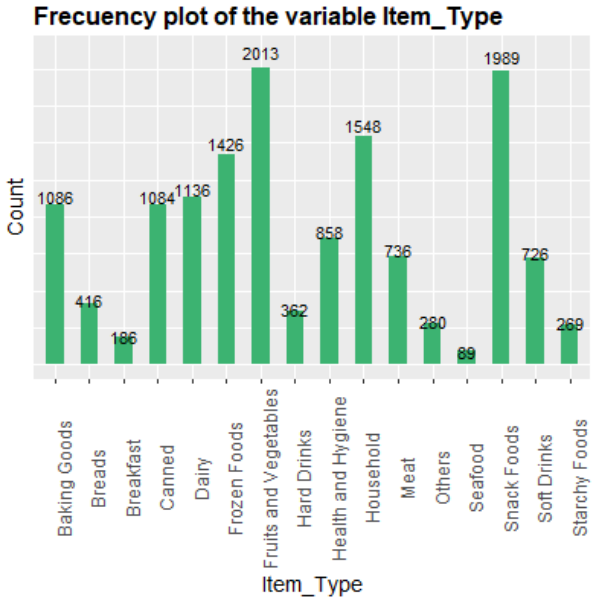
(e) Outlet_Establishment_Year displays a store's age. Comparatively fewer outlets were founded in 1998 than in previous years. The oldest store has opened in 1985 and the

newest one has opened in 2009. During 1999 - 2009 approximately equal number of outlet stores were being opened with supermarket type being the most opened.

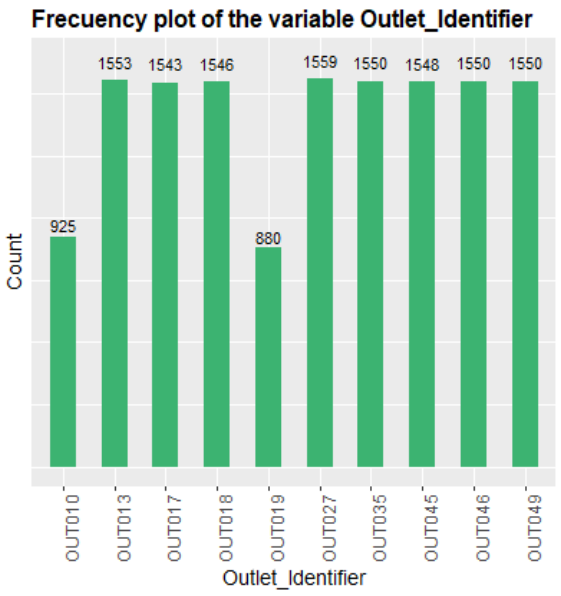
(f) Tire 3 is the highest counted Outlet_Location_Type.



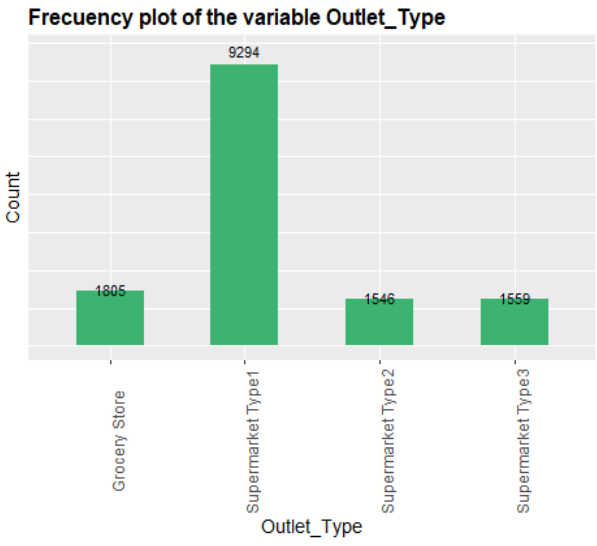
(a)



(b)



(c)



(d)

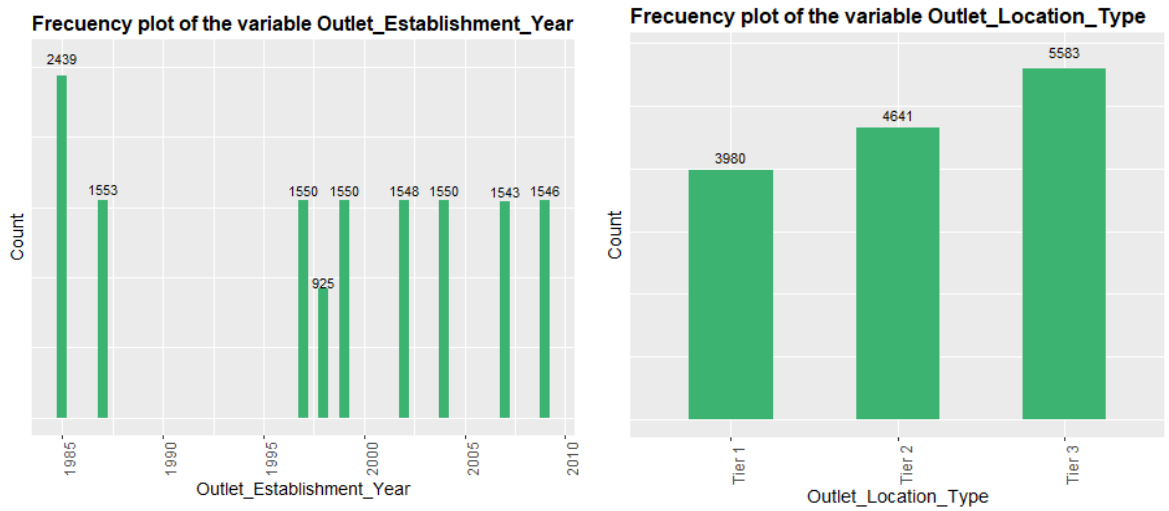
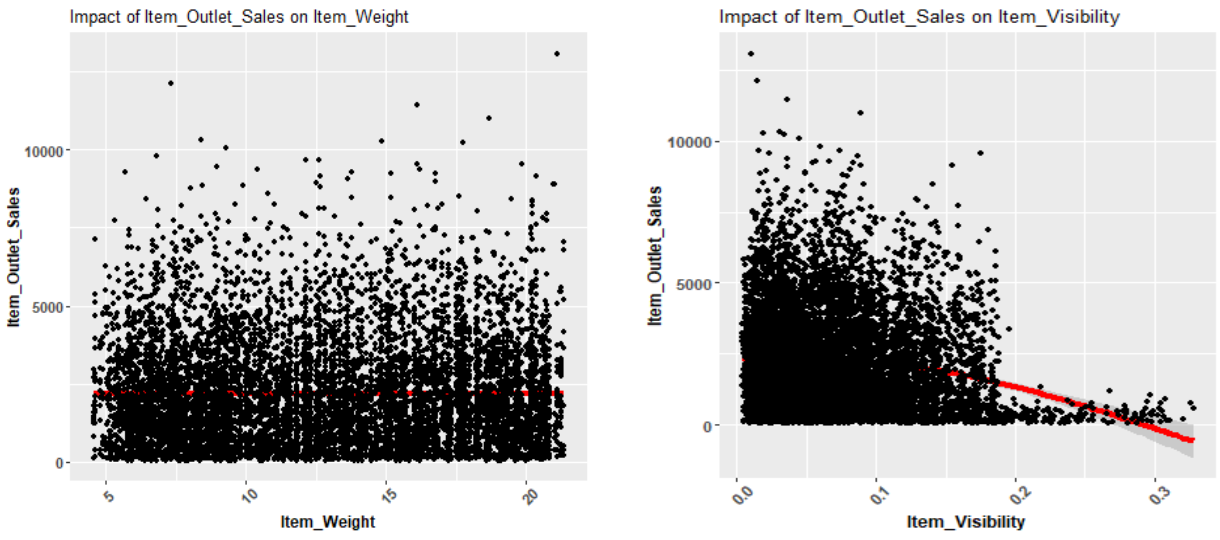


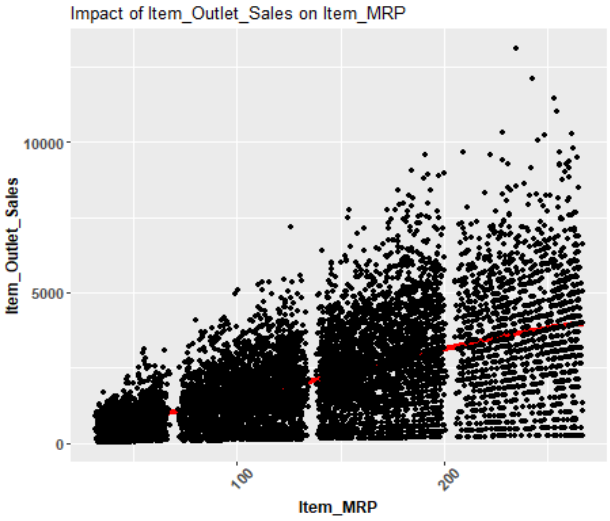
Fig 4.6: Distribution of Independent Categorical Variable (a) Outlet_Size, (b) Item_Type, (c) Outlet_Identifier, (d) Outlet_Type, (e) Outlet_Establishment_Year, (f) Outlet_Location_Type

4.2.2 Bivariate Analysis

4.2.2.1 Impact of Sales Vs Numeric Variables

This visualization shown in figure 4.7 will discover the impact or hidden relationships between the independent numeric variable (Item_Weight, Item_Visibility, Item_MRP) and the target variable (Item_Outlet_Sales) and findings can be used in missing data imputation and feature engineering.





(c)

Fig 4.7: Impact of Item_Outlet_Sales vs Numeric Variables

From figure 4.7 (a), Item_Outlet_Sales is evenly distributed throughout the whole Item_Weight range without exhibiting any particular pattern.

From the relationship graph between Item_Visibility and target variable Item_Outlet_Sales as per figure 4.7(b), less visible items are sold more compared to more visible items. However, the highest dense area of (0.0) creates a dilemma because it suggests that the product is not visible but is sold in the greatest quantity. Additionally, there are things in the outlet that are utilized on a daily basis, which invalidates the null hypothesis.

In the third plot of Item_MRP vs Item_Outlet_Sales from figure 4.7(c), 4 pricing segments are clearly visible that can be used in feature engineering to create a new variable. It is also observed that Items with higher MRP tend to sell better in most cases.

4.2.2.2 Impact of Sales vs Categorical Variables

Now, correlation between categorical variables and Item_Outlet_Sales are visualized grouped by the different Outlet variables like Outlet_Identifier, Outlet_Size, Outlet_Location_Type and Outlet_Type and Outlet_Establishment_Year shown in figure 4.8.

A. Item_Outlet_Sales vs Item_Fat_Content

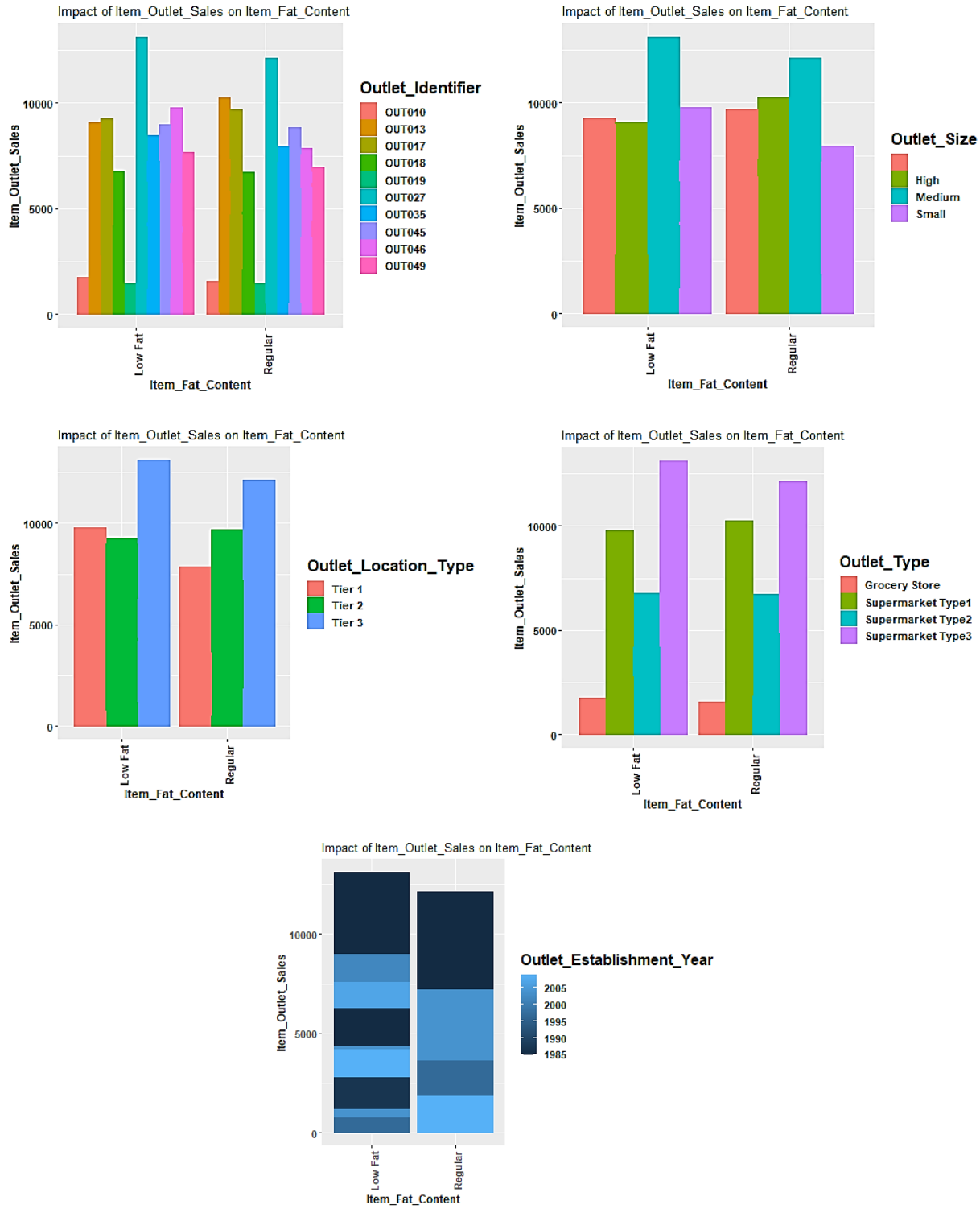


Fig 4.8: Item_Outlet_Sales vs Item_Fat_Content grouped by Outlet variables

It is noticed that Sales of Low Fat Content is high in Medium sized outlet which is located in Tire3 location and Supermarket Type3. Also, the outlets whose establishment year is before 1990 has more Low Fat content sale.

B. Item_Outlet_Sales vs Item_Type

The distribution of Item_Outlet_Sales among the Item_Type categories in the figure 4.9 is not particularly distinct and resembles that of Item_Fat_Content. Sales of Household item is high in every variable category of outlet.

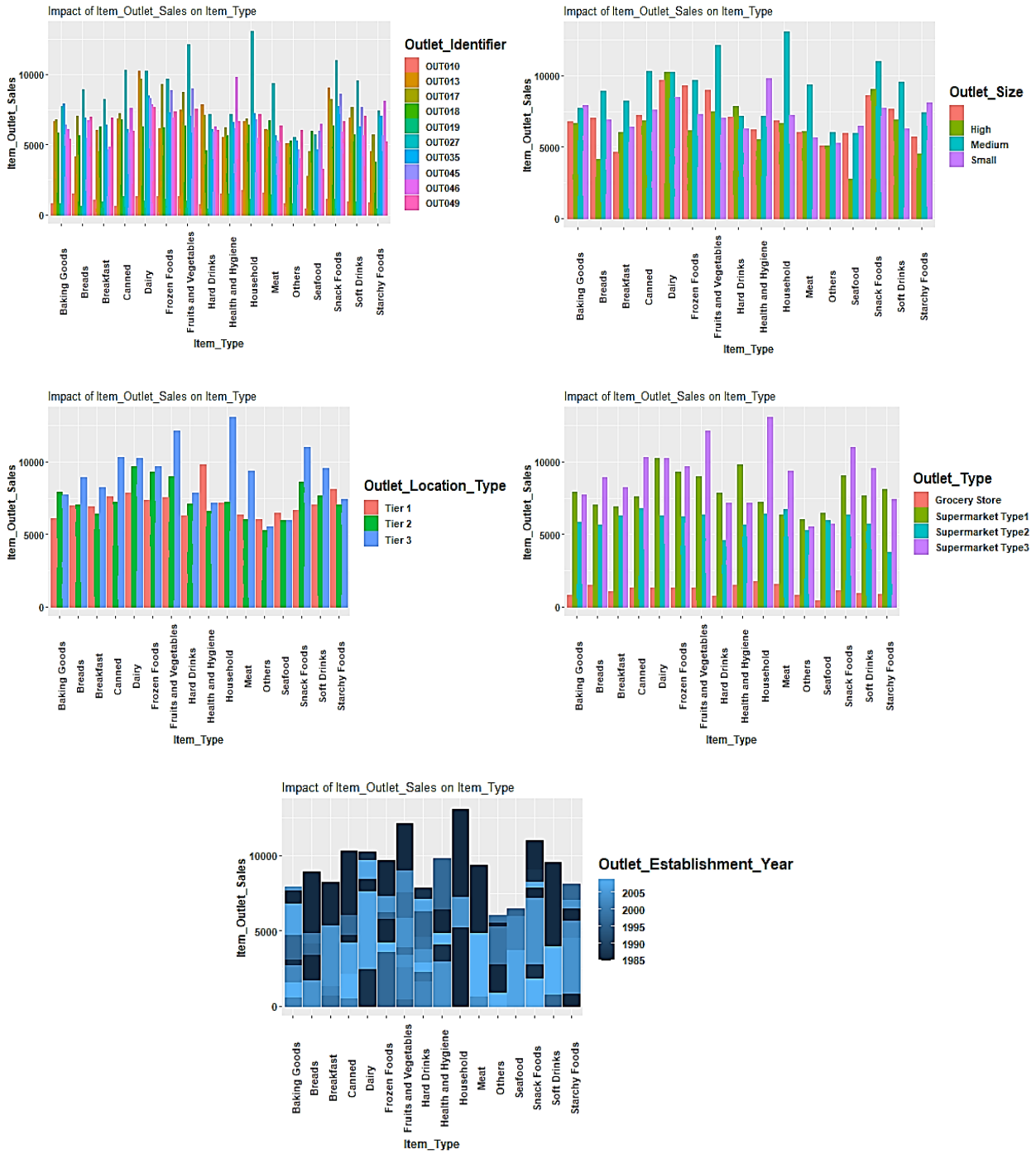


Fig 4.9: Item_Outlet_Sales vs Item_Type grouped by Outlet variables

C. Item_Outlet_Sales vs Outlet_Identifier

From the below plots from figure 4.10, it is clearly found that OUT010 and OUT019 are small in size and the sale of items in these outlets are comparatively very less. Furthermore, the other Outlets having a larger bar plot suggests that the sale of items in these outlets are high whereas some items of Outlet 027 have higher sales when compared to item sales in other outlets. Also Outlet size of OUT027 is medium, located in Tire3 location and in our dataset, only this outlet is Supermarket Type3 and majority of the outlets are Supermarket Type1.

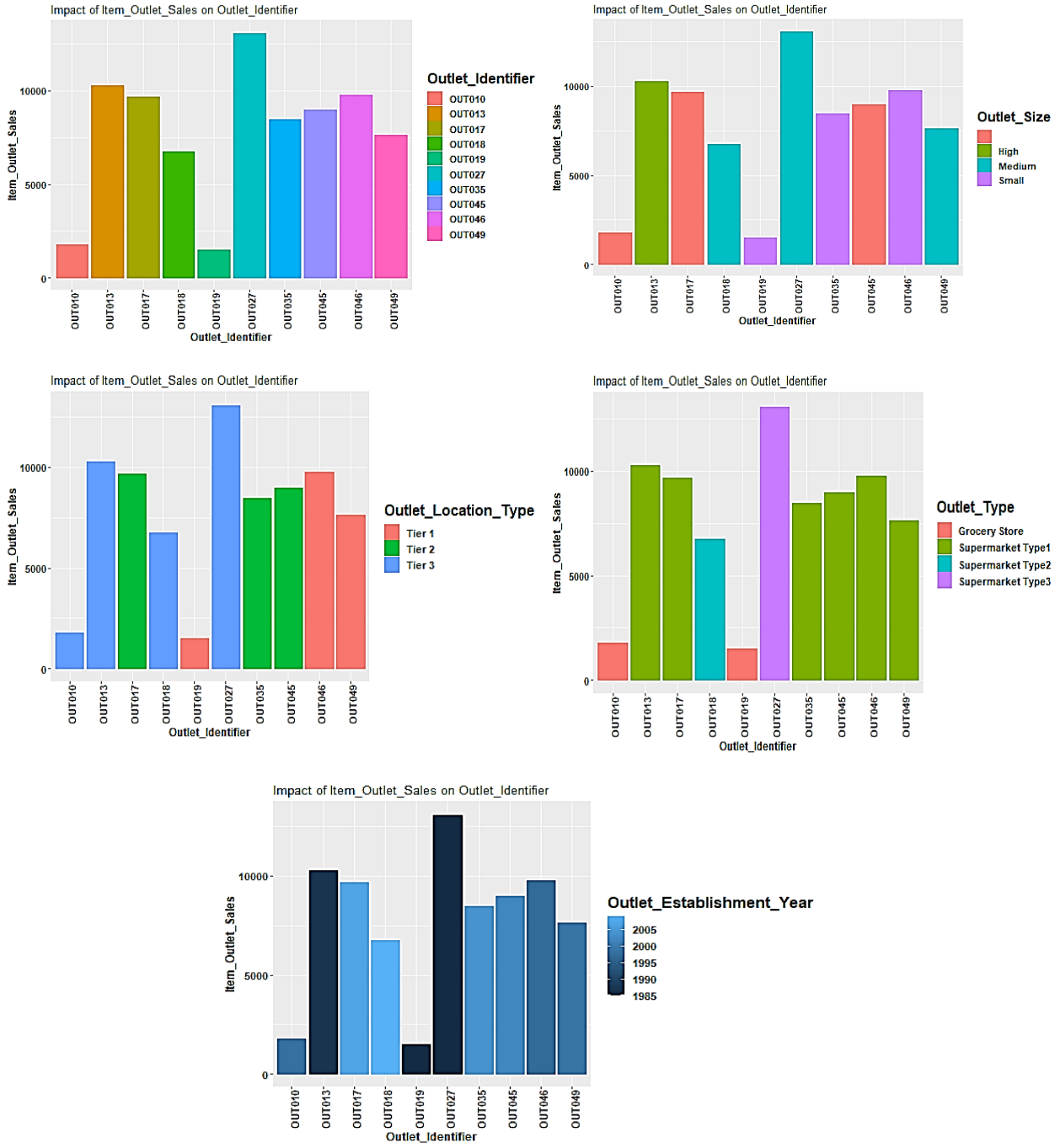


Fig 4.10: Item_Outlet_Sales vs Outlet_Identifier grouped by Outlet variables

D. Item_Outlet_Sales vs Outlet_size

From the below plots in figure 4.11, some missing data of Outlet Size has been found which affects the understanding of sales. Outlet OUT013 is only large sized outlet but not the sales are highest here. Grocery Store Outlet_Type has less sales value items as compared to the other categories. The Medium sized outlets seem to have a larger percentage of high sales items in comparison to the other sizes. Medium and high outlet sizes are pretty much even in sales.

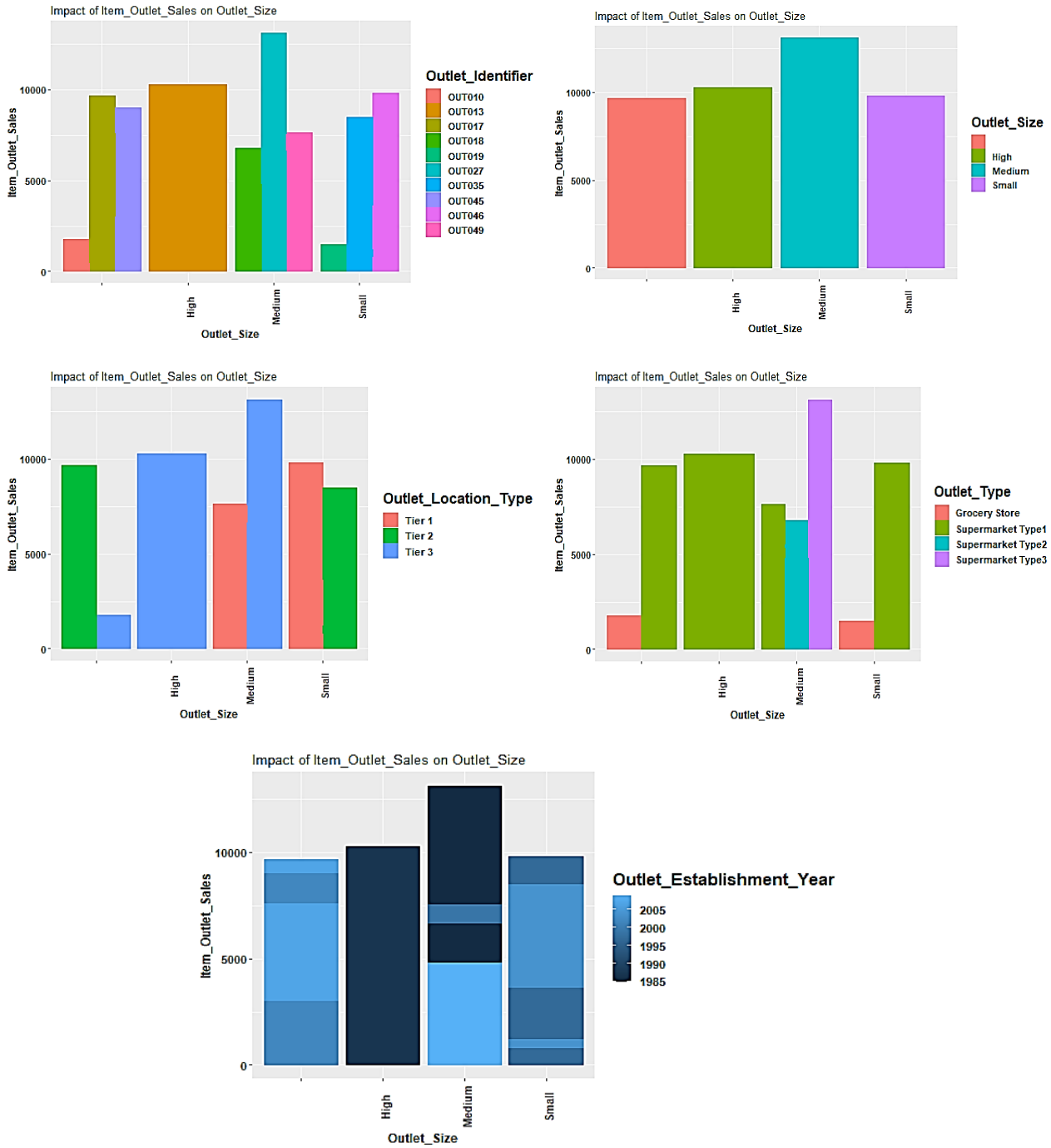


Fig 4.11: Item_Outlet_Sales vs Outlet_Size grouped by Outlet variables

E. Item_Outlet_Sales vs Outlet_Establishment_Year

From the below plots in figure 4.12, the oldest outlet Out027 has the highest impact on the sales.

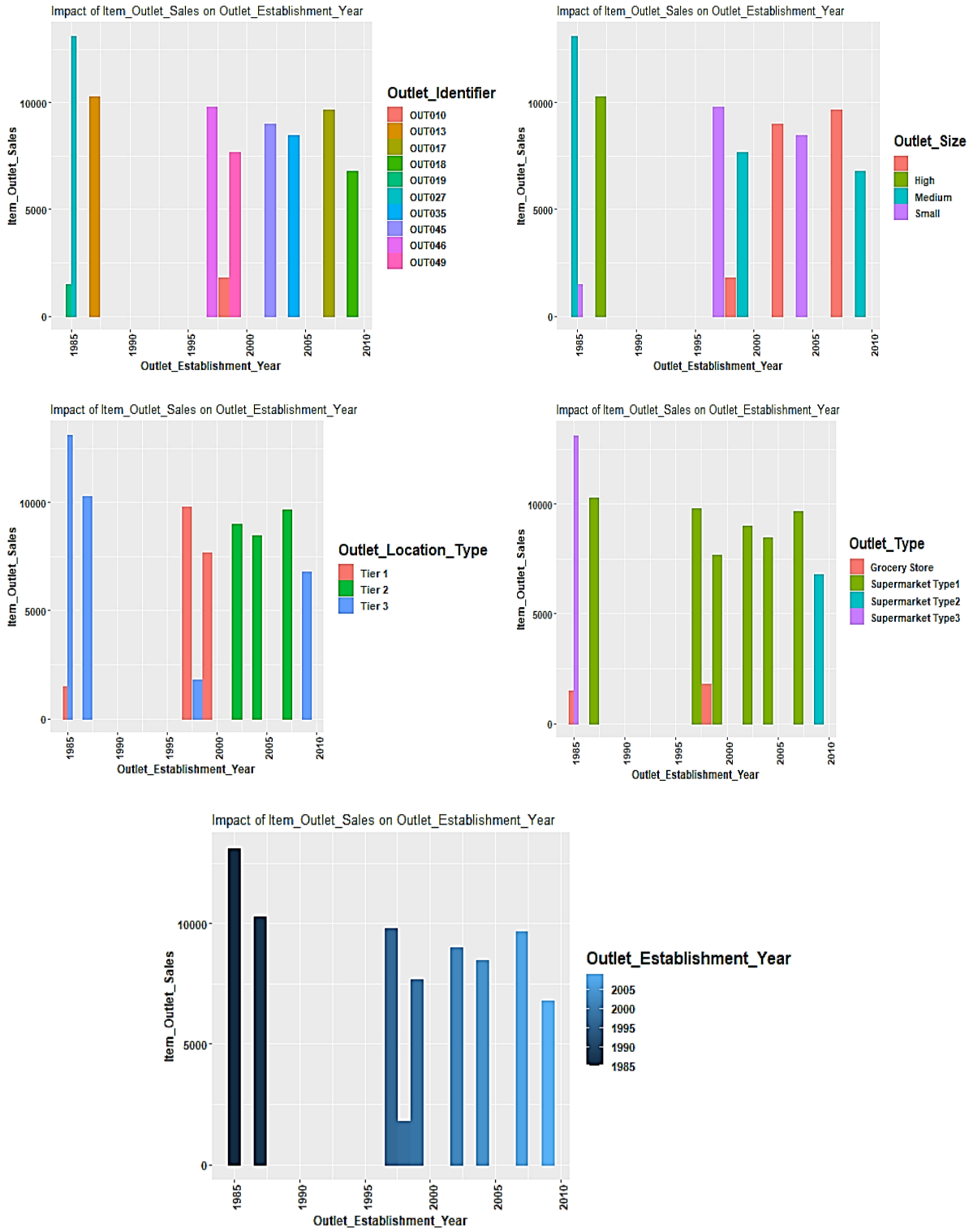


Fig 4.12: Item_Outlet_Sales vs Outlet_Establishment_Year grouped by Outlet variables

E. Item_Outlet_Sales vs Outlet_Type

According to the plots in figure 4.13, Supermarket Type1 is clearly the most frequent form of outlet. However, supermarket type 3 has the biggest sales and is the most profitable.

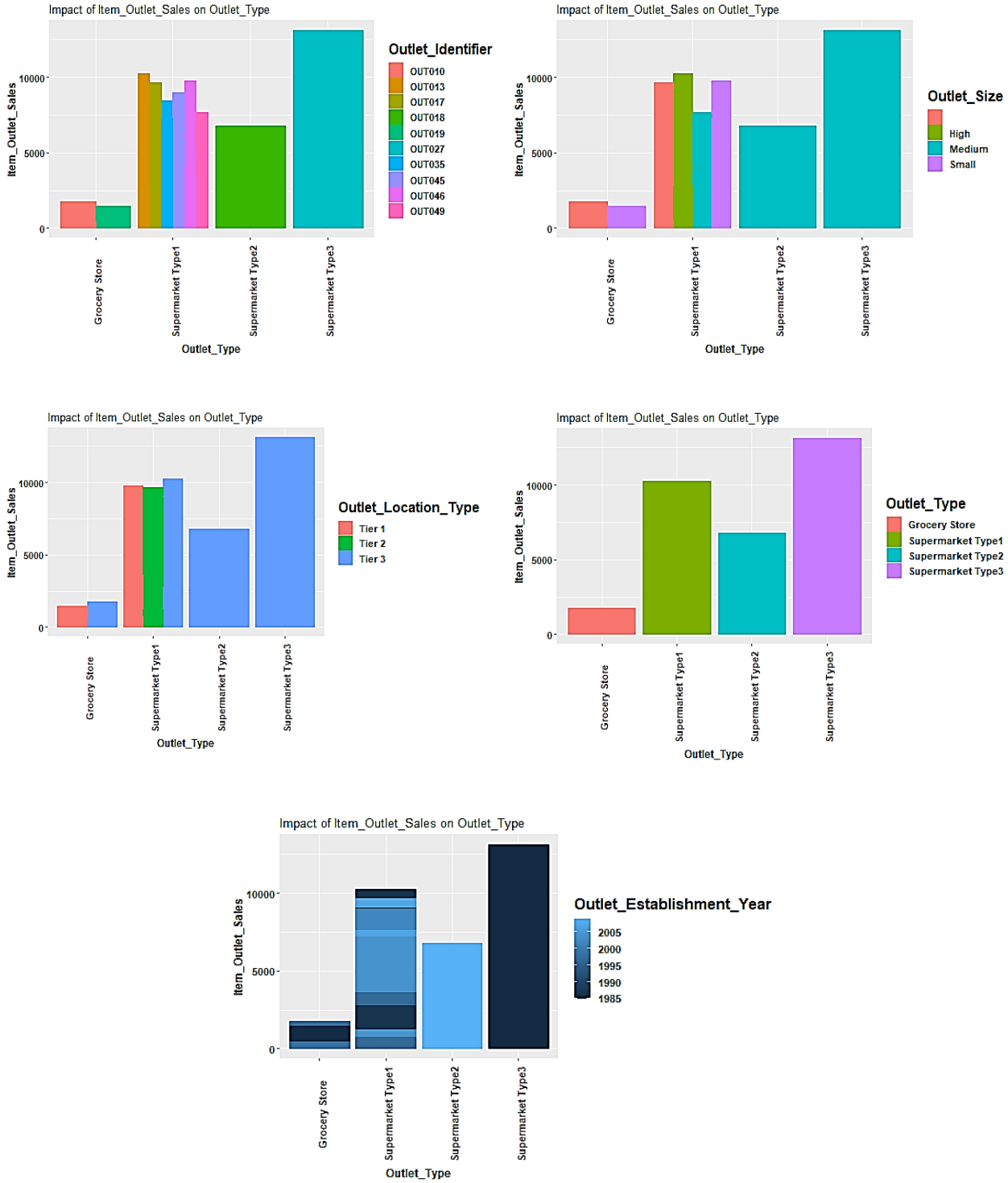


Fig 4.13: Item_Outlet_Sales vs Outlet_Type grouped by Outlet variables

F. Item_Outlet_Sales vs Outlet_Location_Type

From the below figure 4.14, Tier 3 is being the highest in sales whereas Tire 1 and 2 has the quite similar scenario in rest of sales.

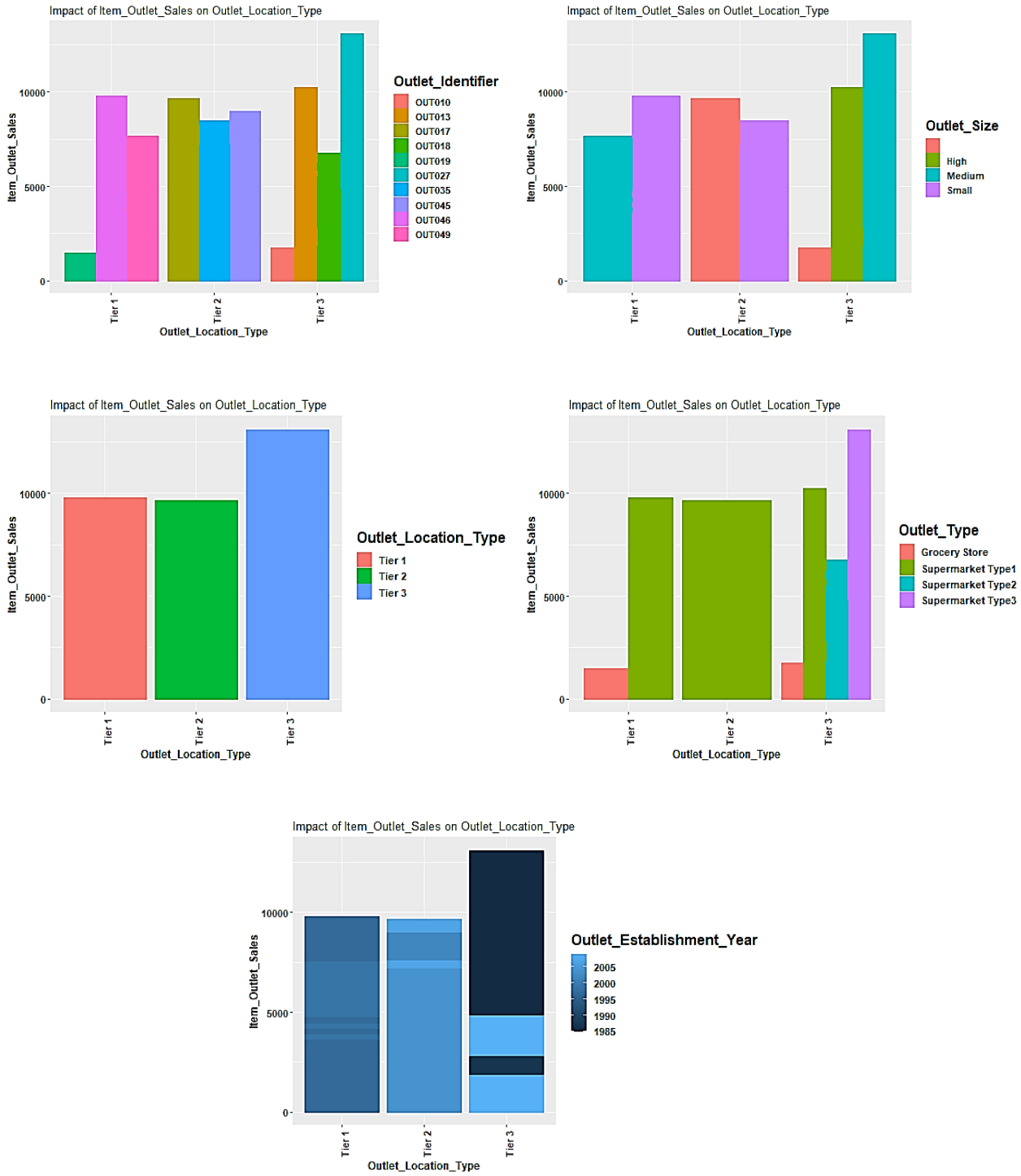


Fig 4.14: Item_Outlet_Sales vs Outlet_Type grouped by Outlet variables

4.3 Data Cleaning

Following the analysis and visualization of the data, the data preparation process must be investigated in light of the conclusions drawn from the data visualization. There is a need to handle any dataset outliers and impute any missing values during this process.

Item_Weight and Outlet_Size, two columns in the dataset, are both missing information for 2439 and 4016 observations, respectively. Outlet_Size is a category variable and Item_Weight is a numeric variable. Therefore, missing values in Item_Weight were imputed with the mean weight based on the Item_Identifier variable as part of the missing value treatment process. The average or mean value cannot be imputed with Outlet_Size. For this procedure, the Outlet_Identifier variable's mode size is utilized. Additionally, the null values in the Item_Visibility field are likewise considered missing values in this instance. Item_Identifier-wise means of the Item_Visibility variable are used in place of the Item_Visibility variable which shown in figure 4.15.

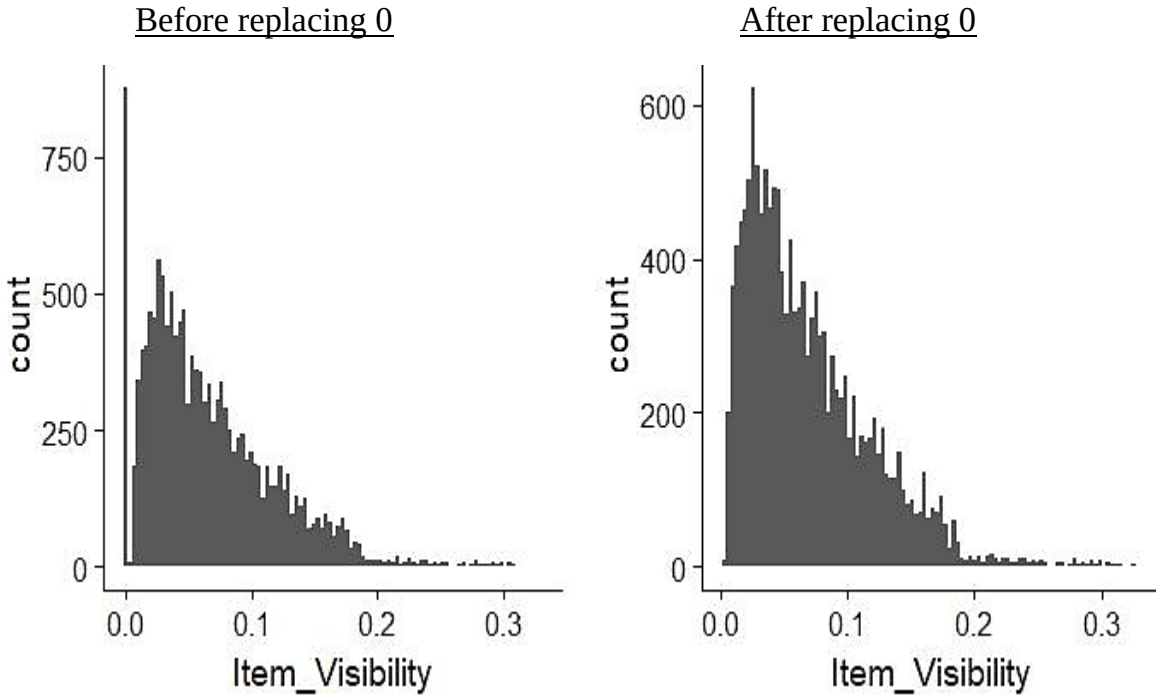


Fig 4.15: The histogram of Item_Visibility before and after replacing zero

4.4 Feature Engineering

In this dataset, From the set of Numeric values, numeric variable Outlet_Establishment_Year has less impact. It is no longer needed. From the set of Categorical values, 7 columns describes in table 4.4 has been found that need to modify or encode:

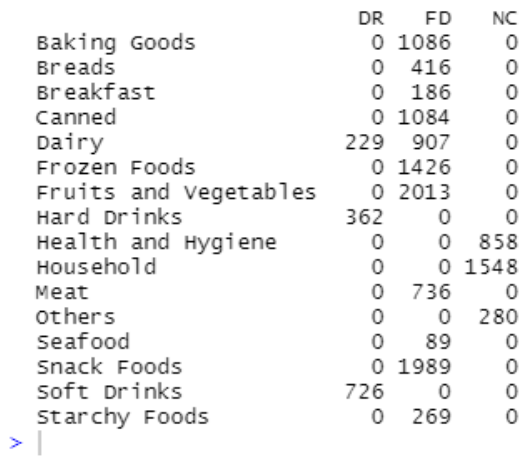
Table 4.4: Feature selection for Feature Engineering

Categorical values		Numeric values
Ordinal variables: **(a)** Item_Fat_Content **(b)** Outlet_Size **(c)** Outlet_Location_Type	Nominal variables: (a) Item_Identifier (b) Item_Type (c) Outlet_Identifier (d) Outlet_Type	(a) Outlet_Establishment_ Year

From the analysis of section 4.2, Outlet_Establishment_Year, Item_Identifier and Outlet_Identifier columns don't have significant values so they will be dropped. Furthermore, all ordinal variables will be Label encoded and Outlet_Type and Item_Type column will be One Hot encoded.

It is observed that the features of given dataset are not enough to give satisfactory prediction. In such scenario, this can be solved by deploying two methods for creating new features such as 1) Hidden Feature extraction, which consists of extracting a new set of features from the already existing features, and 2) External Features, which entails adding more explanatory factors and evaluating their significance and association with our objective variable—the number of sales for each product—while also adding more explanatory variables. External Features, which is to add more explanatory variables and checking their importance and correlation with our target variable, being number of sales for each product. The four categories of new features that has been developed for the first methodology are as follows:

(a) **Item_category.** Item_Identifier was used to create the categorical variable. It was noted that each unique ID in the Item_Identifier property began with either FD, DR, or NC. So, a new column called "Item_category" was added, with the categories "Foods," "Drinks," and "Non-consumables."



	DR	FD	NC
Baking Goods	0	1086	0
Breads	0	416	0
Breakfast	0	186	0
Canned	0	1084	0
Dairy	229	907	0
Frozen Foods	0	1426	0
Fruits and Vegetables	0	2013	0
Hard Drinks	362	0	0
Health and Hygiene	0	0	858
Household	0	0	1548
Meat	0	736	0
Others	0	0	280
Seafood	0	89	0
Snack Foods	0	1989	0
Soft Drinks	726	0	0
Starchy Foods	0	269	0

>

Fig 4.16: Item Categories

(b) **Outlet_Years.** Years of operation for outlets for which the dataset includes a new column Outlet_Years that identifies how old a specific outlet is.

(c) **Item_Type_new.** Broader classifications for the Item_Type variable. In order to create a new feature, categories will be classified into perishable and non _perishable using the Item_Type variable.

(d) **price_per_unit_wt.** One more new column named price_per_unit_wt is added with the value of Item_MRP/Item_Weight to find out the price per unit weight.

Since non-consumable items in the Item_Category "NC" cannot contain any fat, there is a requirement to change the values of Item_Fat_Content.

The second approach involves removing all data rows with null values and adding additional attributes known as explanatory variables to the data set. It is seen that the Item's MRP was divided into 4 chunks from the plots of Item_MRP vs. Item_Outlet_Sales. As a result, Item_MRP variable can be used to create 4 groups using k Means clustering. Prior knowledge of K, or the number of clusters intend to divide data into, is necessary for K Means clustering. Here the process will proceed with K=4 and add the following column to the dataset:

(a) **Item_MRP_clusters:** Histogram feature for Item_MRP.

4.5 Correlation

The correlation graphic that follows displays correlation between every possible pair of variable and its potential pair in the processed dataset. In order to find hidden patterns among variables, R program corrplot is being used here, which offers a visual exploring tool on

correlation matrices and permits automatic variable reordering. In the corrplot package, there are 7 visualization methods (parameter methods): circle, square, ellipse, number, shade, color, and pie. The correlation between any two variables in this project is shown by a colored square. Positive correlation is represented by a blueish square, whereas negative correlation is represented by a redish square. The square's colored area indicates the correlation's strength. For the purpose of deciding the next course of action, the correlation plot between numerous independent and dependent variables is analyzed.

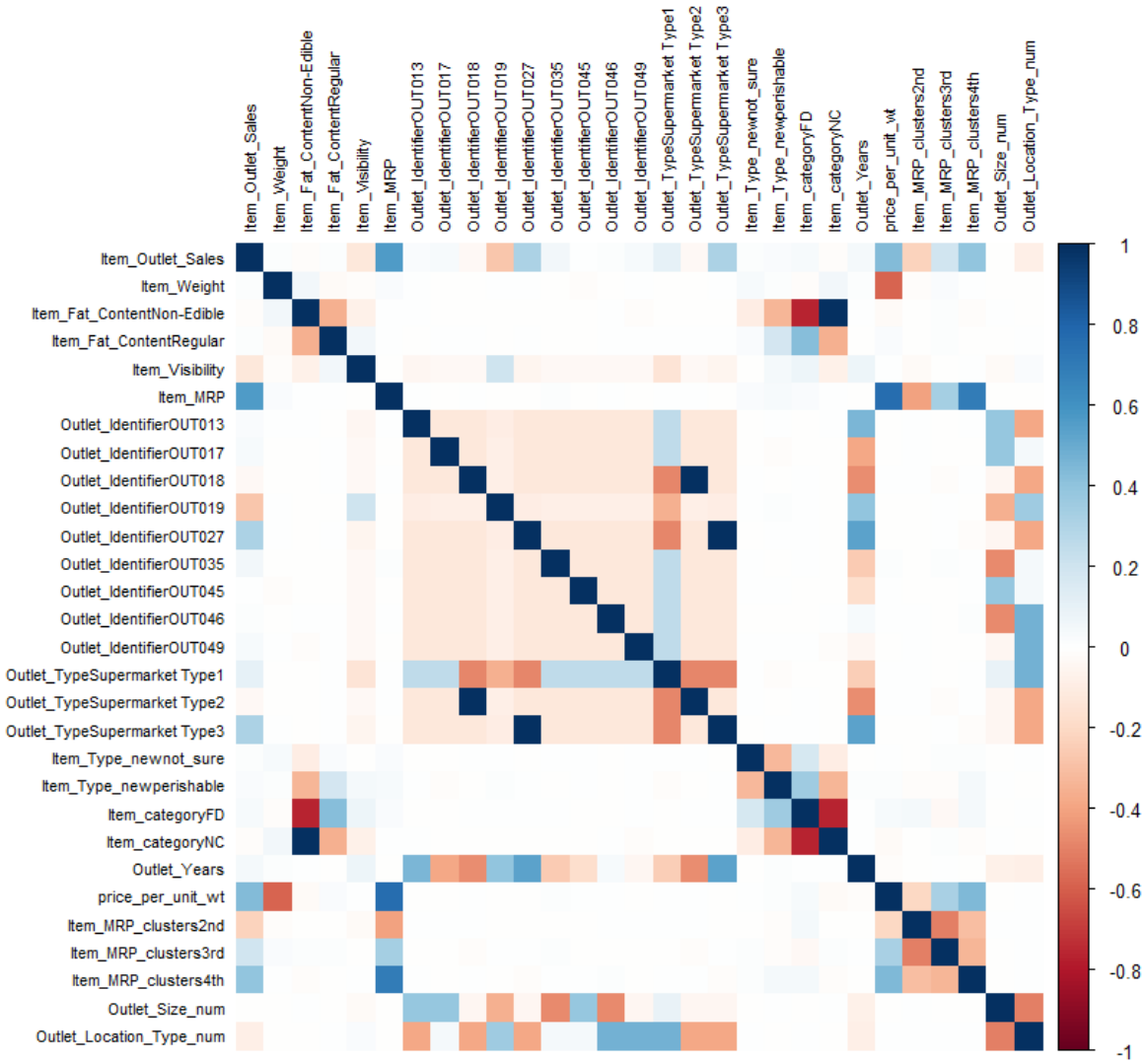


Fig 4.17: Diagram showing correlation among different factors

According to the correlation plot created using the processed dataset, a growing number of variables are displaying their correlation, and the following are some of the key findings about some of the employed variables:

a. Because there is a negative association between Item_Visibility and Item_Outlet_Sales, increasing Item_Visibility can lower Item_Outlet_Sales. This suggests that Item_Visibility does not affect sales, which runs counter to the widespread belief that things with greater visibility generate higher sales.

b. There is a positive correlation between an item's MRP (maximum retail price) and its in-store sales. Therefore, Item_MRP may be a crucial factor in estimating Item_Outlet_Sales at a specific store.

c. The Outlet Type Supermarket_Type3 has good sales and a positive correlation as well.

d. Considering that price_per_unit_wt is derived from the Item_Weight variable, their correlations are very strong. For the same reason, price_per_unit_wt and Item_MRP have a strong correlation.

4.6 Predictive Model Building

After several iterations of data preparation and processing, the appropriate dataset has been obtained on which to develop a predictive model. Upon partitioning data into training and training sets, the models were trained for each methods and evaluate their effectiveness with the evaluation matrices. Later on, applying the predictive model on our test data. Storing the result from our test data in a separate file with a predicted Item_outlet_sales column. For this process, different packages of R have been used for easy deployment of algorithms on both train and test dataset. This will serve as the regression model for this project because this project is predicting numerical values. Following machine learning models deployed here on training dataset and the model were also tested for performance analysis.

(a) Linear Regression

(b) Random Forest

(c) Decision tree

(d) XGBoost Regressor

These chosen algorithms employ various model-training strategies. In order to evaluate the effectiveness of various approaches for the use of sales forecasting, these methodologies have been carefully chosen. According to previously published work that has been reviewed, these chosen models are also used frequently in this research.

For the models constructed in this project, 5-fold cross validation has been applied. A model's generalizability to new data is revealed. Furthermore, R packages and libraries such as xgboost, rpart, randomforest was used for model building. Also, for training the models, method ‘lm’ for linear regression, ‘rpart’ for decision tree, ‘ranger’ for random forest model has been used. For XGBoost, the xgb.cv () function is used which comes with the XGBoost package.

CHAPTER 5
MODEL EVALUATION AND RESULT ANALYSIS

5.1 Model Evaluation

As briefly mentioned in the previous chapter, Linear Regression, Random Forest, Decision Tree, and XGBoost Regressor are trained with the set of data using a 5-fold cross-validation approach that dynamically selects the training and testing with fixed proportion each time. The efficiency was determined using RMSE, MAE, and Rsquared value using cross validation methods.

5.1.1 Linear Regression

The model summary in figure 5.1 shows the resampling of dataset using 5 fold cross-validation and the results of RMSE, Rsquared and MAE values.

```
Linear Regression

8523 samples
  28 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6819, 6818, 6819, 6817, 6819
Resampling results:

  RMSE      Rsquared   MAE
  1130.643  0.5612935  838.8781

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Fig 5.1: Model Summary of Linear Regression

Additionally, the diagnostic plot in Figure 5.2 displays a scatterplot of the residual prediction errors against the projected values and is used to see whether the predictions may be made more accurate by addressing issues with our data.

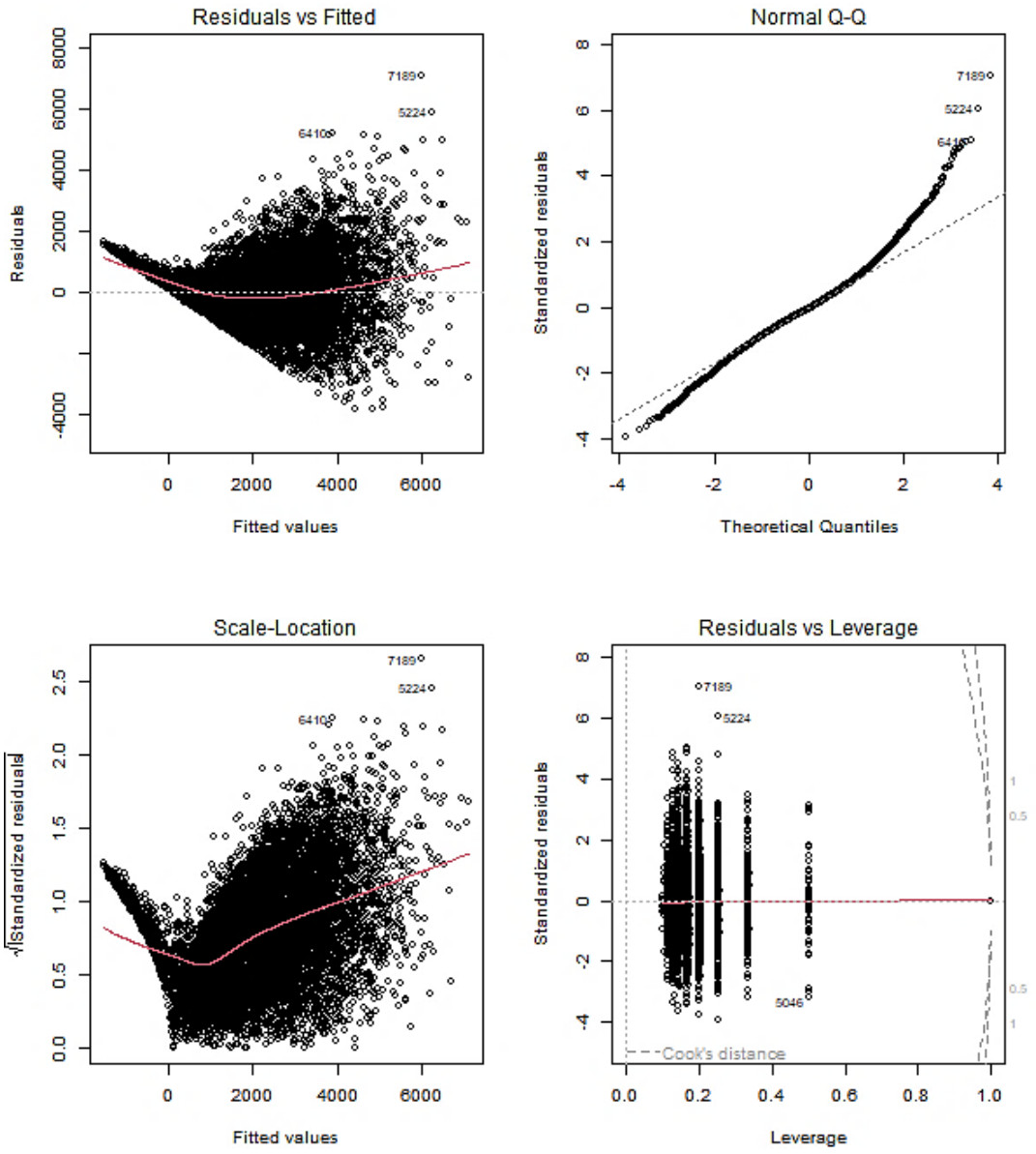


Fig 5.2: Diagnostic plot of Linear Regression

5.1.2 Decision Tree

The model summary in figure 5.3 shows the resampling results across the tuning parameter, complexity parameter (cp) using 5-fold cross-validation and the optimal value used for this model is cp=0.01 which makes the improved RMSE value over the linear regression model. Figure 5.4 shows the Cross-Validation RMSE plot for decision tree model.

```
CART

8523 samples
  28 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6819, 6818, 6818, 6817, 6820
Resampling results across tuning parameters:

  cp    RMSE      Rsquared   MAE
  0.01  1123.941  0.5662304  812.3749
  0.02  1175.541  0.5254404  863.0273
  0.03  1189.462  0.5141792  873.2985
  0.04  1253.303  0.4606696  947.4696
  0.05  1253.303  0.4606696  947.4696

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.01.
```

Fig 5.3: Model Summary of Decision Tree

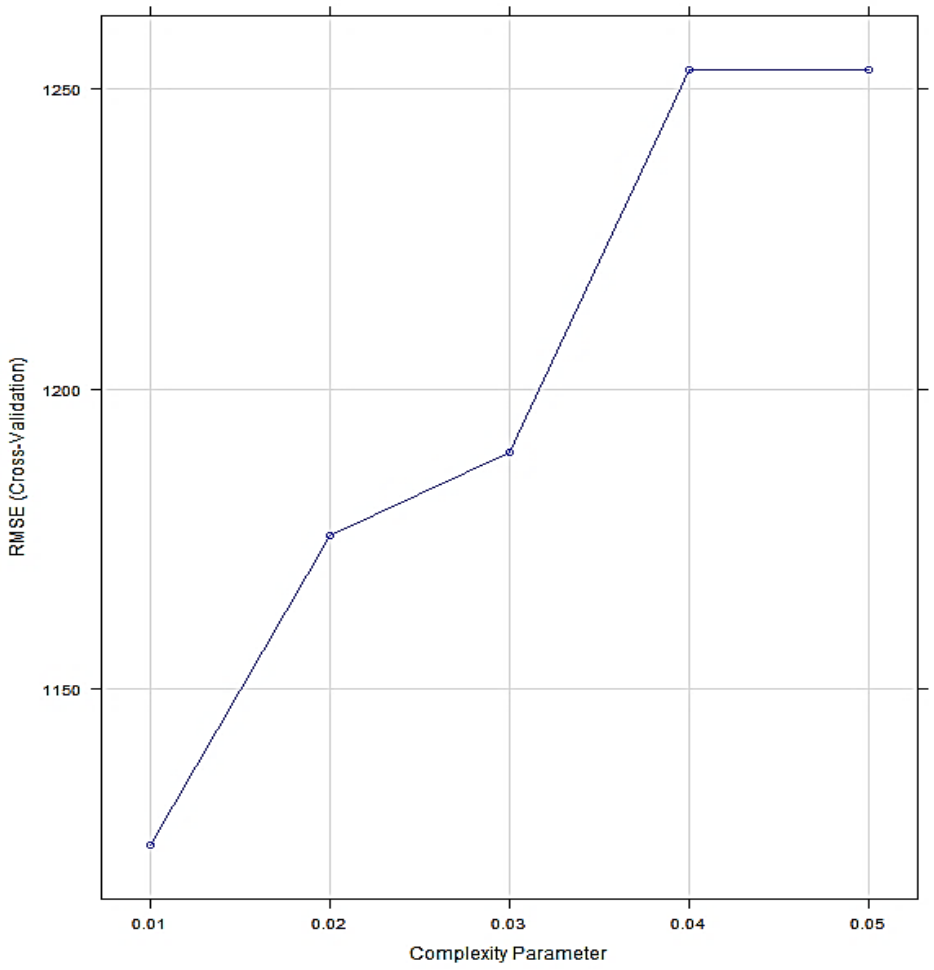


Fig 5.4: RMSE (Cross Validation) for Decision Tree

5.1.3 Random Forest

The model summary in figure 5.5 shows the resampling results across the tuning parameter, mtry and min.node.size using 5 fold cross-validation and the optimal value used for this model is no. of predictor variables, mtry = 5 and minimum size of terminal nodes, min.node.size = 20 which makes the improved RMSE value over the decision tree model. Figure 5.4 shows the Cross-Validation RMSE plot for Random Forest model.

```
Random Forest

8523 samples
  28 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 6819, 6817, 6819, 6818, 6819
Resampling results across tuning parameters:

  mtry  min.node.size  RMSE      Rsquared   MAE
   3    10             1109.264  0.5884140  801.2680
   3    15             1109.325  0.5885325  801.7418
   3    20             1109.783  0.5887137  803.3355
   4    10             1090.832  0.5930715  772.4753
   4    15             1091.878  0.5925293  774.3639
   4    20             1091.765  0.5928910  774.4373
   5    10             1088.296  0.5933063  766.2647
   5    15             1087.040  0.5943275  765.5906
   5    20             1086.932  0.5945099  765.7309
   6    10             1093.650  0.5891792  768.3198
   6    15             1089.765  0.5920405  766.0950
   6    20             1087.823  0.5935184  764.1151
   7    10             1095.982  0.5875931  769.6076
   7    15             1091.522  0.5908322  766.6763
   7    20             1090.507  0.5915586  766.2638
   8    10             1097.569  0.5866182  770.3507
   8    15             1094.643  0.5886887  768.1281
   8    20             1091.695  0.5907995  766.3099
   9    10             1101.525  0.5838949  773.2589
   9    15             1095.881  0.5878588  769.0782
   9    20             1093.964  0.5892125  767.9094
  10    10             1101.265  0.5842323  772.2580
  10    15             1097.986  0.5864724  770.3406
  10    20             1094.692  0.5887841  768.0393

Tuning parameter 'splitrule' was held constant at a value of variance
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mtry = 5, splitrule = variance and min.node.size = 20.
```

Fig 5.5: Model Summary of Random Forest

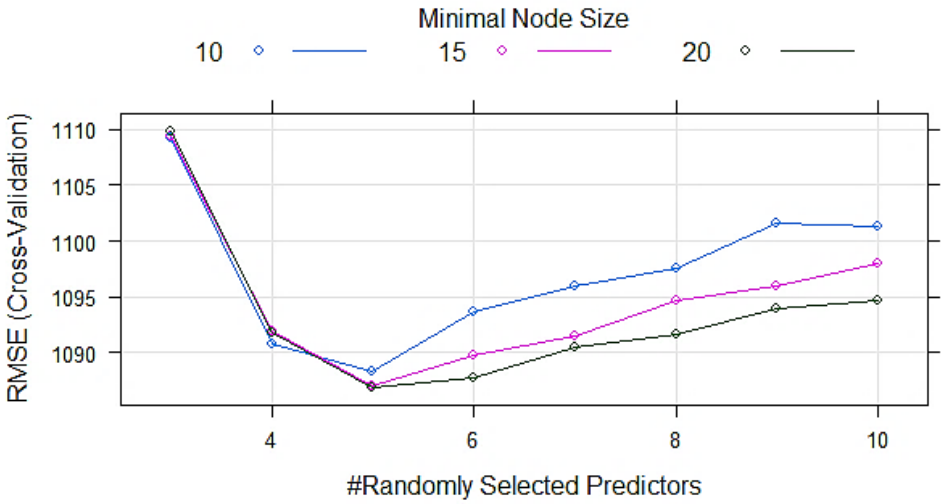


Fig 5.6: RMSE (Cross Validation) for Random Forest

5.1.4 XGBoost

From the model of XGBoost, the best iteration has been found on iter=683, train-rmse:1014.882, train-mae:722.0682; test-rmse:1083.657, test-mae:764.046 which is representing in figure 5.7 and 5.8.

```
##### xgb.cv 5-folds
    iter train_rmse_mean train_rmse_std train_mae_mean train_mae_std test_rmse_mean test_rmse_std te
       1        2747.020      10.750748      2159.0212      7.478494       2746.725      46.79328
       2        2724.916      11.234745      2137.4341      7.432766       2724.602      46.20748
       3        2704.426      11.657024      2116.0991      7.369307       2704.012      45.91353
       4        2683.285      12.275032      2095.0274      7.330589       2682.888      45.21424
       5        2662.596      12.567052      2074.1928      7.226108       2662.153      44.63857
---
     709        1012.470       7.396433       720.5884      5.600059       1083.951      26.49411
     710        1012.392       7.414164       720.5367      5.607010       1083.957      26.53396
     711        1012.304       7.416810       720.4836      5.605654       1083.978      26.51329
     712        1012.220       7.410702       720.4317      5.603756       1083.976      26.51650
     713        1012.150       7.420950       720.3798      5.611911       1083.960      26.50566
    test_mae_std
        30.04487
        29.99399
        30.07284
        30.06148
        30.02350
---
        19.10137
        19.13163
        19.11094
        19.12319
        19.09950
Best iteration:
 iter train_rmse_mean train_rmse_std train_mae_mean train_mae_std test_rmse_mean test_rmse_std test_
  683        1014.882       7.437059       722.0683      5.629688       1083.657       26.8016
 test_mae_std
     19.21788
```

Fig 5.7: Model Summary of XGBoost

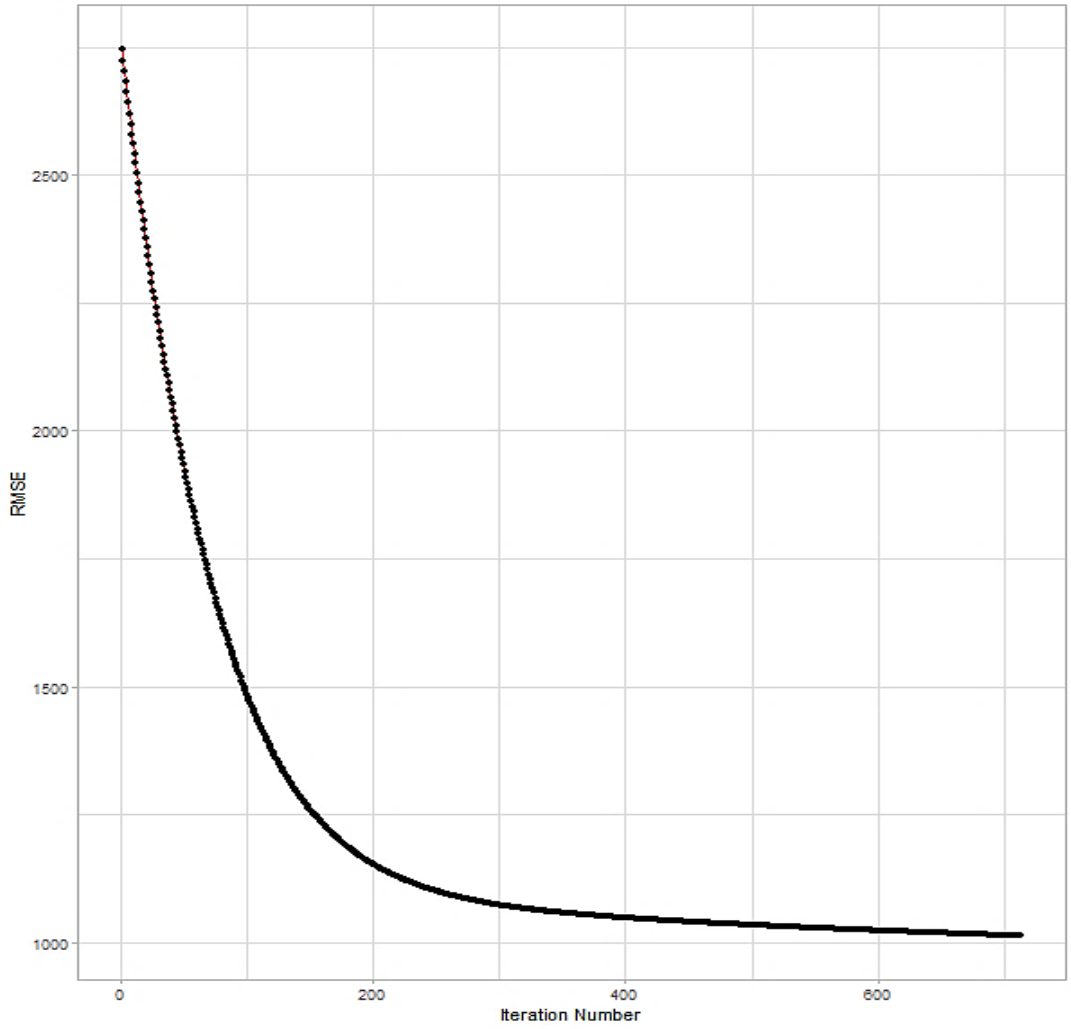


Fig 5.8: RMSE (Cross Validation) for XGBoost

5.2 Result Analysis

5.2.1 Result Analysis with Evaluation Metrics

According to the findings, from the deployment of models it has been seen that Random forest model have the highest Rsquared value which is 0.594 whereas Rsquared value for Decision tree is 0.566 and linear regression model is 0.561. From this performance analysis, Random Forest found as better model than Decision tree and Linear Regression. Table 5.1 shows the comparison of Rsquared (R^2):

Table 5.1: Performance Analysis with Rsquared value

Model Name	Rsquared (R^2)
Linear Regression	0.561
Decision Tree	0.566
Random Forest	0.594

Results from Figures. 5.1, 5.3, 5.5, and 5.7 demonstrate the RMSE and MAE of the prediction models that the algorithms produced. The MAE scores for the Linear Regression, Decision Tree, Random Forest and XGBoost were 838.87, 812.37, 765.73, 722.068 and RMSE score 1130.64, 1123.94, 1086.93, 1014.88 respectively. It is observed that, XGBoost Model has the lower RMSE and MAE value than other models. Comparing evaluation results, Table 5.2 demonstrates that XGBoost performed satisfactorily for all metrics tested, including RMSE and MAE. When compared to Linear Regression, Decision Tree, and Random Forest, XGBoost had the lowest error in forecasting the sales. The following table 5.2 illustrates the comparative analyses of the implemented algorithm based on prediction performance:

Table 5.2: Performance Analysis with RMSE and MAE

Model Name	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)
Linear Regression	1130.64	838.87
Decision Tree	1123.94	812.37
Random Forest	1086.93	765.73
XGBoost	1014.88	722.068

5.2.2 Result Analysis with Feature Importance

The model evaluation results have led us to the conclusion that Item_MRP is the most crucial variable in predicting the target variable, which is a common finding across all models. Except for Item_MRP variable, each model lists a separate set of attributes as critical features.

According to Figure 5.9, the decision tree model's estimation of the pricing feature relevance for the products Item_MRP, Item_MRP_clusters3rd, and Item_MRP_clusters2nd would be mostly based on sales.

```
variable importance
                  Item_MRP         Item_MRP_clusters3rd
                        18                           10
     Item_MRP_clusters2nd            price_per_unit_wt
                        10                            9
             Outlet_Years         Item_MRP_clusters4th
                         9                            6
  Outlet_IdentifierOUT018 Outlet_TypeSupermarket Type2
                         6                            6
  Outlet_IdentifierOUT027 Outlet_TypeSupermarket Type3
                         6                            6
Outlet_TypeSupermarket Type1     Outlet_Location_Type_num
                         6                            3
  Outlet_IdentifierOUT019             Outlet_Size_num
                         3                            2
```

Fig 5.9: Feature Importance from Decision Tree

Figure 5.10 depicts the Random Forest model's feature importance, with Item_MRP as the standout feature because it is significantly dependent on sales and the other characteristics don't even come close. These characteristics will undoubtedly have a significant impact on sales forecasting.

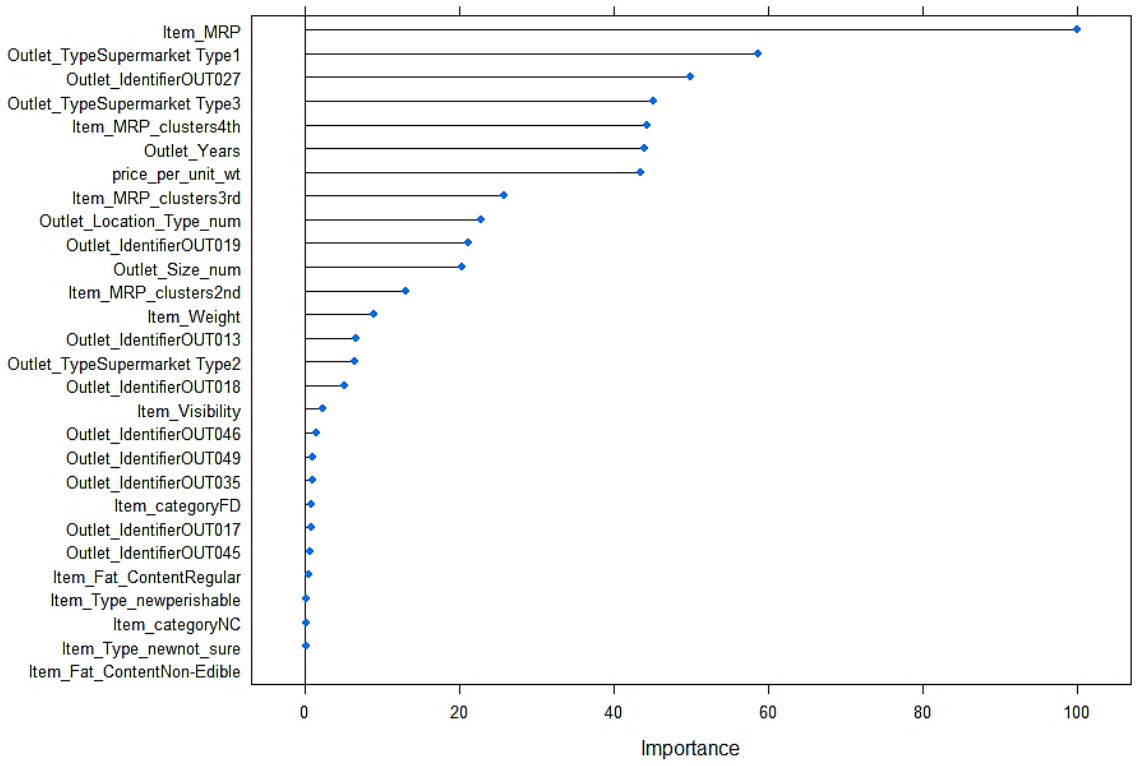


Fig 5.10: Feature Importance plot from Random Forest Model

Additionally, Item_MRP is the most crucial variable in the XGBoost model depicted in figure 5.11. New features like price_per_unit_wt, Outlet_Years, and Item_MRP_Clusters are also among the top factors to consider.

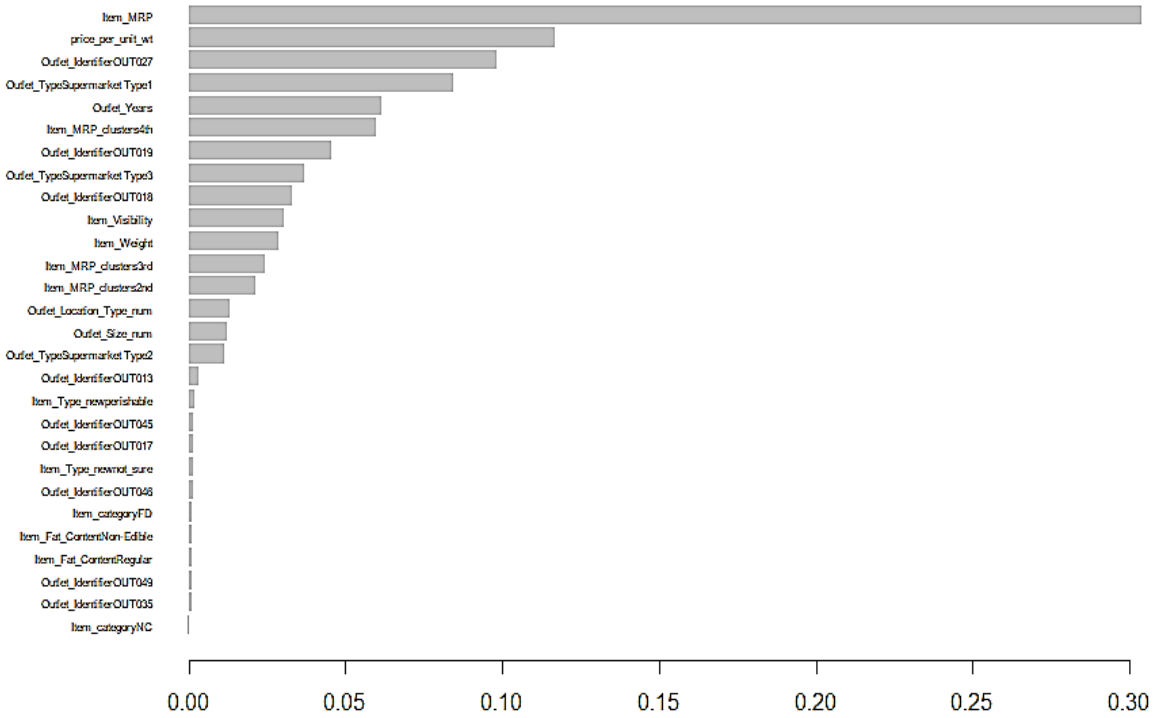


Fig 5.11: Feature Importance plot from XGBoost

5.3 Discussion

The goal of this study is to evaluate how well data mining and machine learning methods predict the sales of particular items at outlets using a collected data through a software prototype for the industry usage. For the purpose of this study, the dataset was "cleaned" using some data preprocessing techniques. When the dataset was verified for missing values, it was discovered that one column had a sizable number of missing values; these missing values were then replaced with the column's mean and mode of size. The dataset used includes both category and numerical input variables; in order to make the most of the algorithms, the categorical variables were transformed into numerical variables using the R library's dummy variable package. To ensure a thorough learning process, the feature scaling technique was applied to the complete set of data. In an effort to identify patterns, trends, traits, and anything else that might be of interest, data exploration was carried out on the dataset. The study has found that the larger location didn't generate the highest sales, according to the data exploration.

When compared to other outlets, the outlet known as "outlet 027" had greater item sales. Some things in the outlet were sold more than the other items in the outlet. Additionally, the Supermarket Type3, one of the outlet types, demonstrates that the sale of goods in that group is higher than that of other groups and that there is a greater difference in sales amongst the goods. Additionally, compared to outlets founded in previous years, those formed in 1985 saw the highest item sales and the greatest discrepancy in item sales. This study showed that the various outlets, outlet types, and the year that the outlets were founded all play a significant part in the forecasting of outlet item sales. The models used to forecast the sales of outlet items were developed using machine learning techniques. Mean Square Error (MSE) and Root Mean Square Error were used to assess the performance of the machine learning methods used to train the predictive models (RMSE). The XGBoost Regression model outperformed the other machine learning methods in terms of performance. From the software prototype solution, any user can visualize the data analysis portion with the graphical representation for the similar dataset. Additionally, they can also use the model for knowing the prediction results of Item_Outlet_Sales from the collected dataset where it was unknown to them earlier.

CHAPTER 6
DEPLOYMENT OF APPLICATION SOFTWARE

6.1 Architecture of Application Software dashboard

In order to determine the optimal approach for incorporating machine learning approach in to problem statement that are used in the industry, a deployment plan is being used in this project. By integrating with the applied machine learning models, a prototype of application software is being developed.

In this project R Shiny Dashboard is used as part of the deployment of application software prototype as a web service. Figure 6.1 shows the design architecture of the sales prediction dashboard mentioning the different modules from which the existing featured can be visualized at a glance.

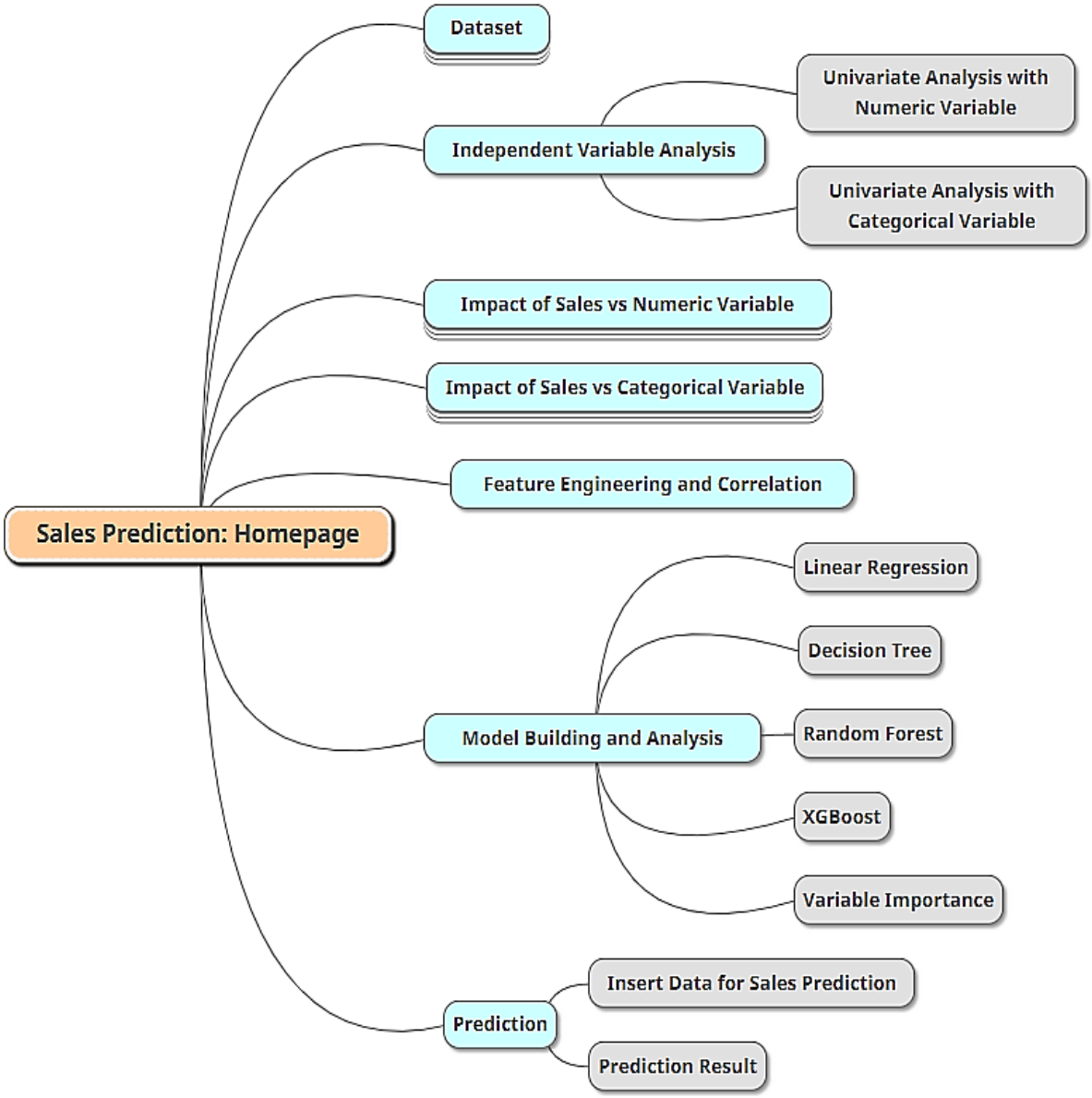


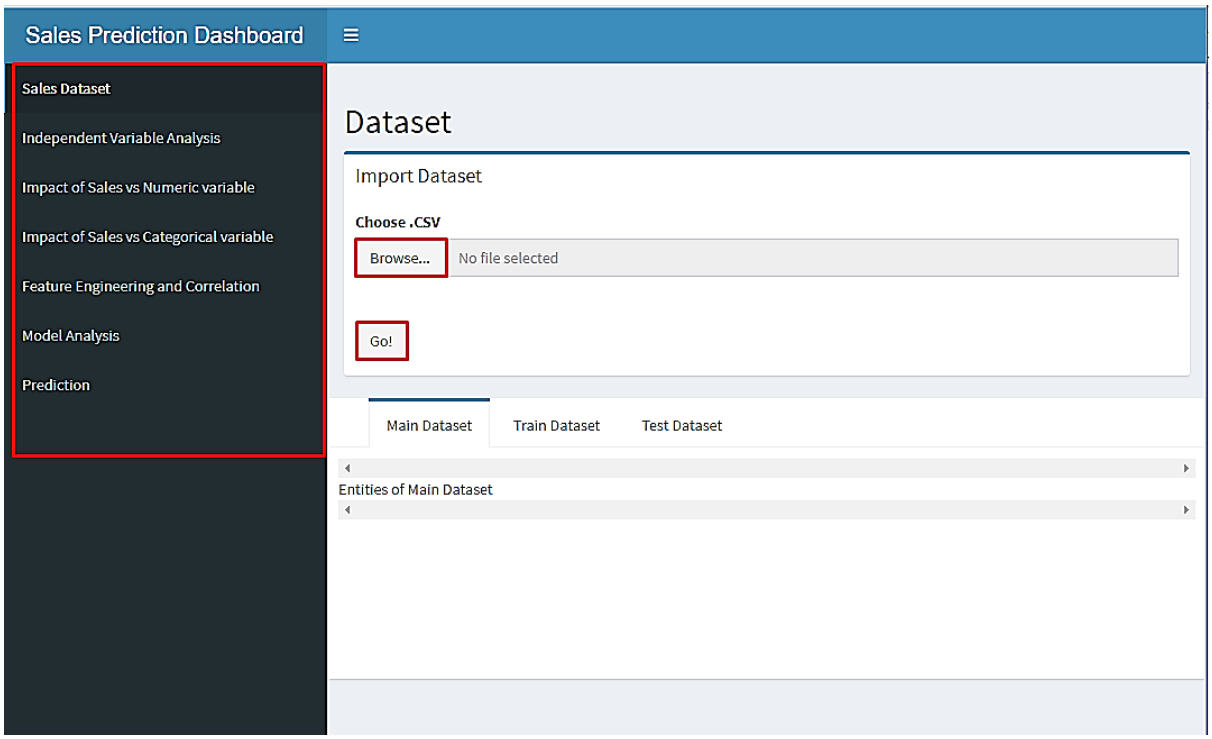
Fig 6.1: Architecture of Sales Prediction Dashboard

6.2 Sales Prediction Dashboard Overview

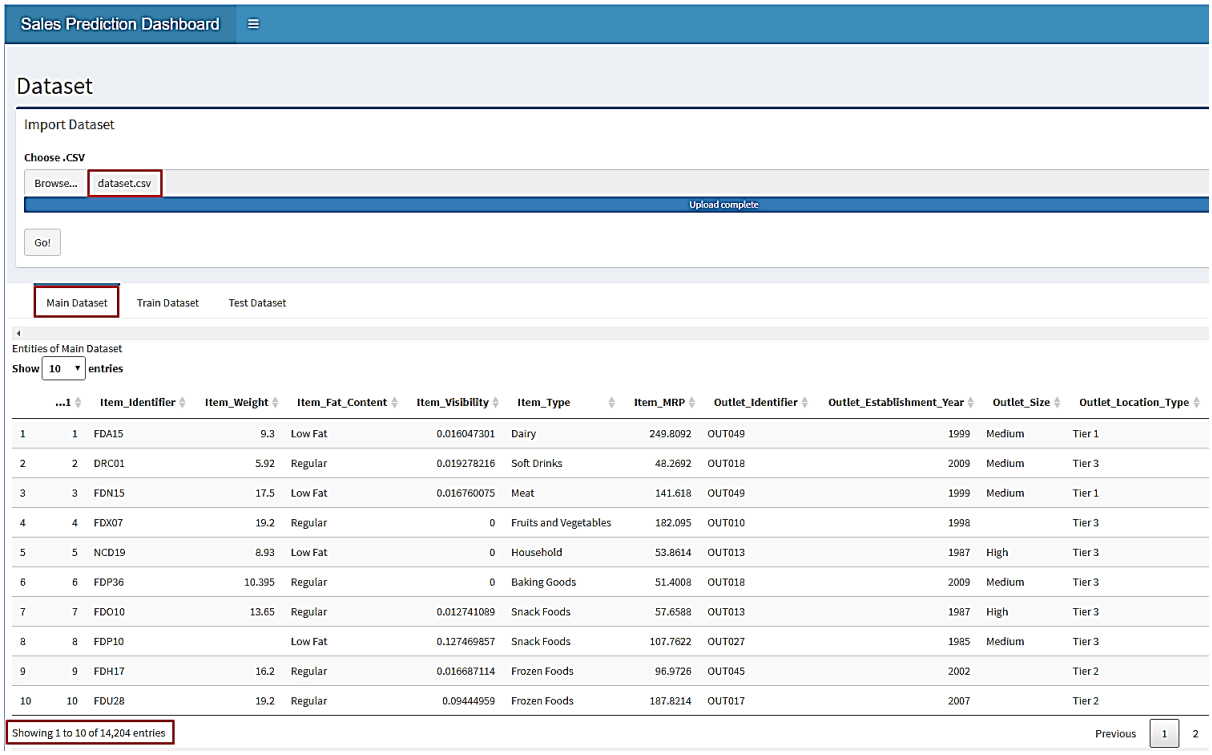
6.2.1 MenuItem-1: Sales Dataset

From the Sales Prediction Dashboard, the side menu bar contains several menu-items shown in the left side of figure 6.2(a). In this figure, Sales Dataset side-menu bar contains the tab view of main dataset, train dataset and test dataset. Here, the feature of importing .CSV file as a dataset has been integrated. It helps a user to import similar dataset as shown in figure 6.2(b)

to make the prediction and understand the dataset with impact of target variable or relationship with target variable.



(a)



(b)

Fig 6.2: Sales Prediction Dashboard (a) Opening preview of Sales Prediction Dashboard, (b) Dashboard view after importing Sales Dataset

6.2.2 MenuItem-2: Independent Variable Analysis

From the Sales Prediction Dashboard, Independent Variable Analysis side-menu bar contains the graphical representation of numerical variables and categorical variables which can be selected using Drop-down selection as shown in below figure 6.3.

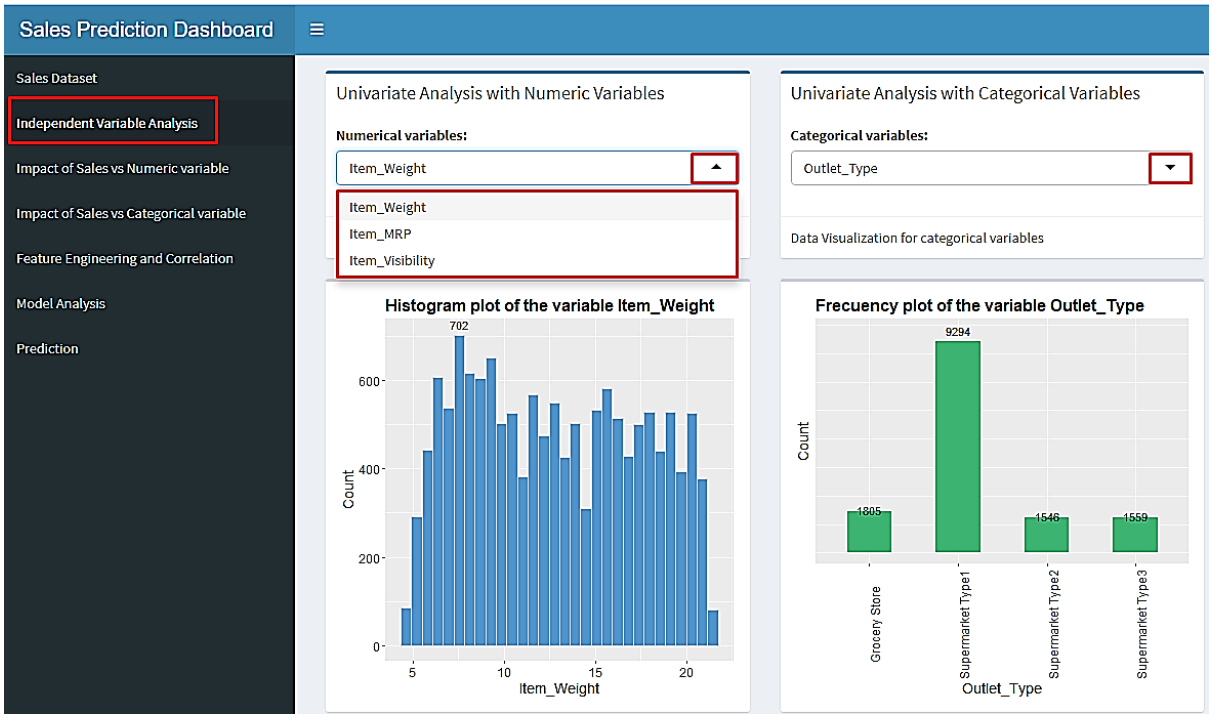


Fig 6.3: Dashboard view of Independent Variable Analysis

6.2.3 MenuItem-3: Impact of Sales vs Numeric Variables

From the figure 6.4 of Sales Prediction Dashboard, Impact of Sales vs Numeric Variables side-menu bar contains the graphical representation of Bivariate analysis of numerical variables in respect to target variable Item_Outlet_Sales where variables can be selected using Drop-down selection.

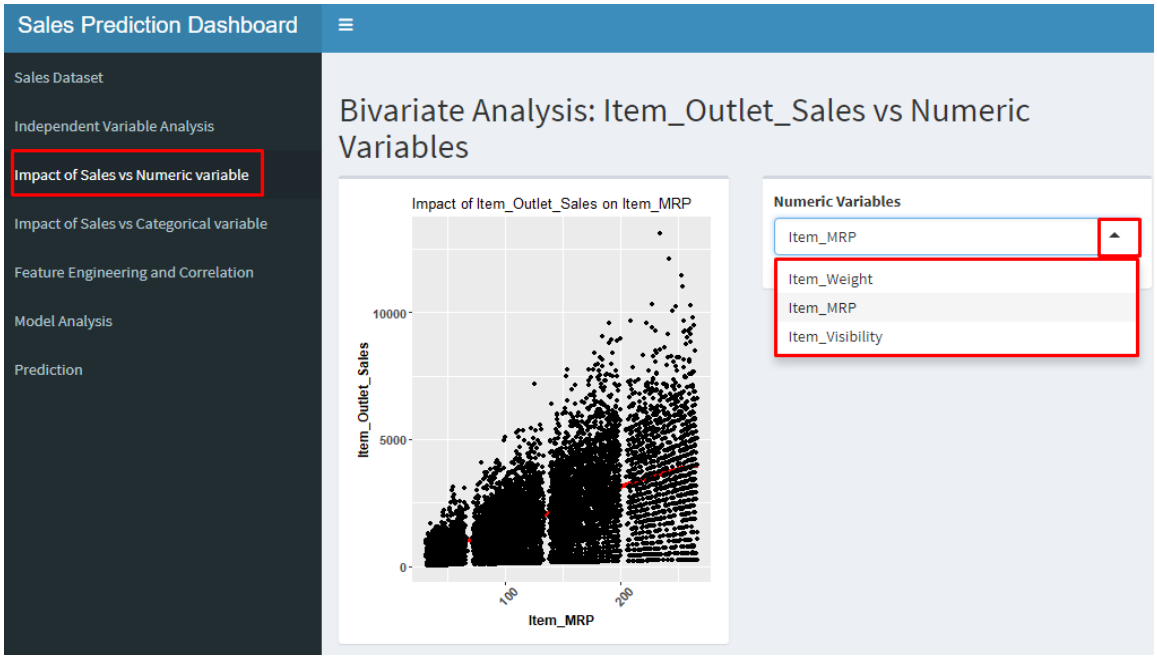


Fig 6.4: Dashboard view of Impact of Sales vs Numeric Variables

6.2.4 MenuItem-4: Impact of Sales vs Categorical Variables

Impact of Sales vs Categorical Variables side-menu bar shown in figure 6.5 which contains the graphical representation of Bivariate analysis of categorical variables in respect to target variable Item_Outlet_Sales grouped by Outlet variables where categorical and outlet variables can be selected using Drop-down selection.

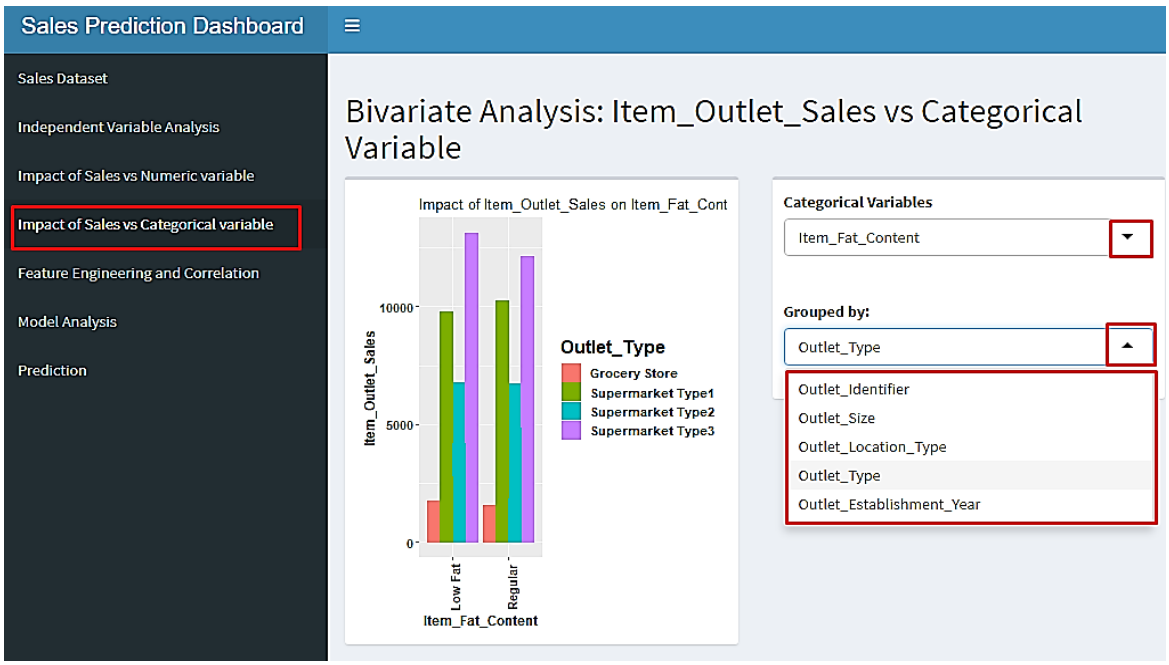


Fig 6.5: Dashboard view of Impact of Sales vs Categorical Variables

6.2.5 MenuItem-5: Feature Engineering and Correlation

Feature Engineering and Correlation side-menu bar shown in figure 5.17 contains the graphical view of correlation between every possible pair of variable and its potential pair in the processed dataset.

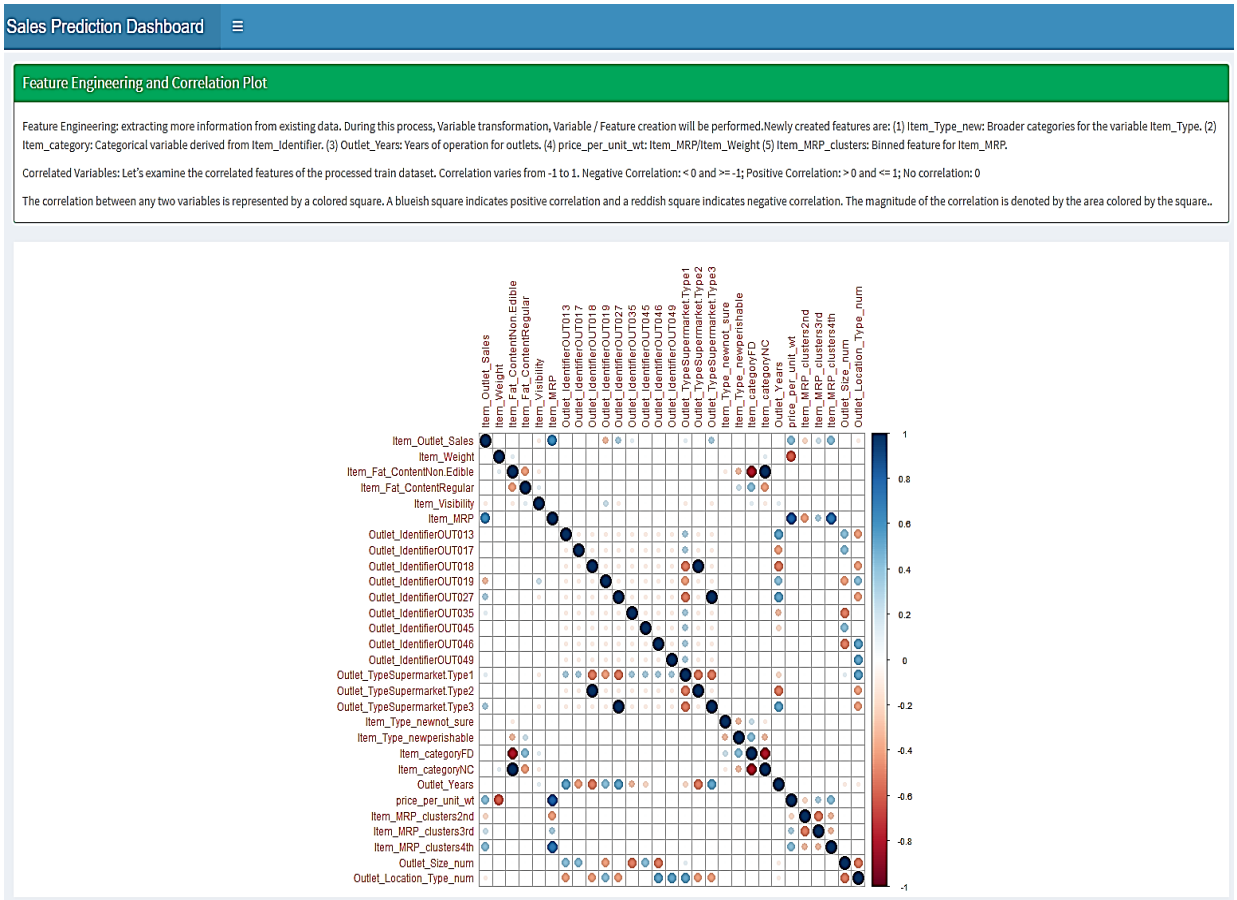


Fig 6.6: Dashboard view of Feature Engineering and Correlation

6.2.6 MenuItem-6: Model Analysis

Model Analysis side-menu bar contains the detailed analysis representing the plots and summary of each models as shown in figure 6.7, 6.8, 6.9 and 6.10.

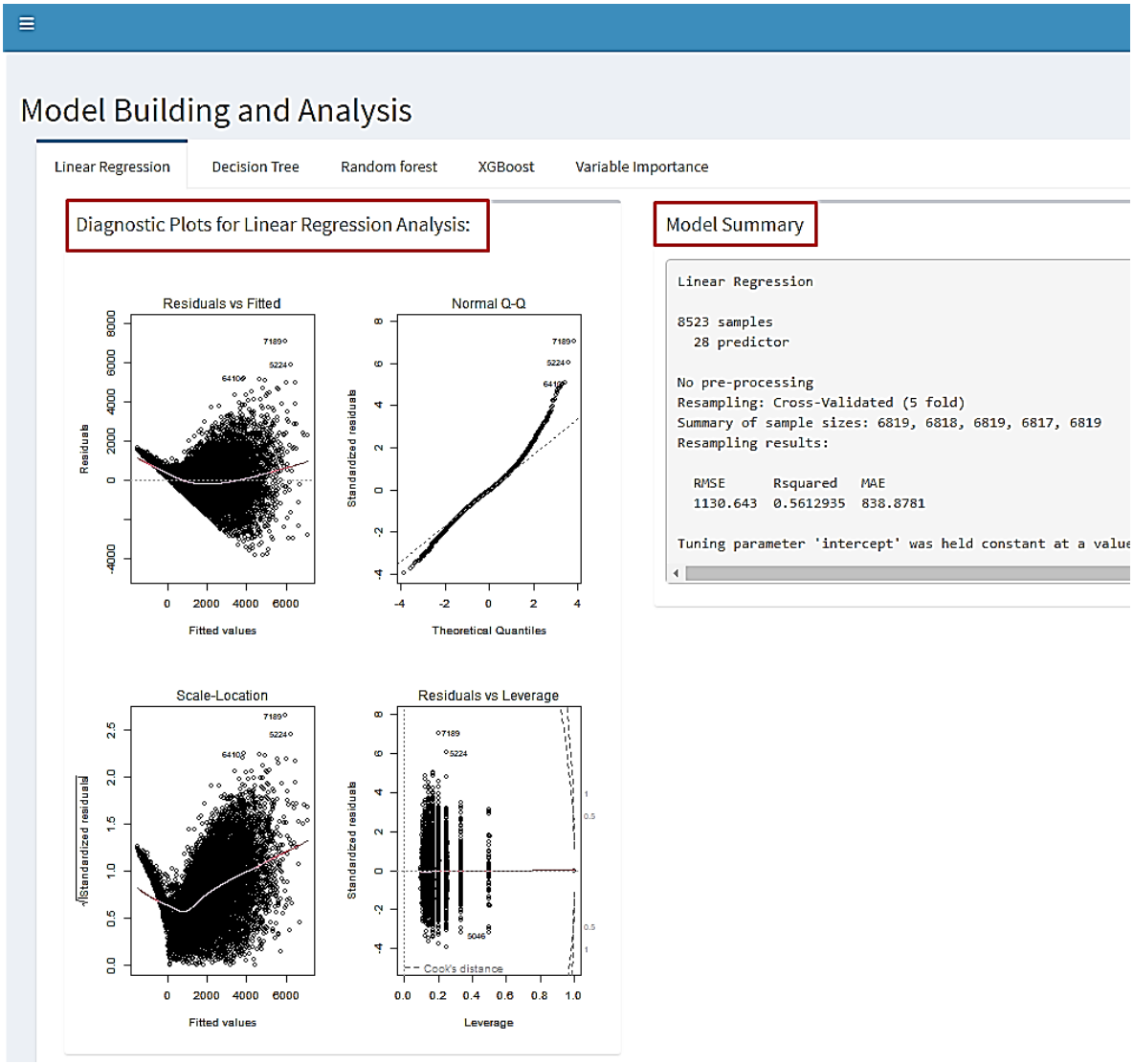


Fig 6.7: Linear Regression Model Analysis

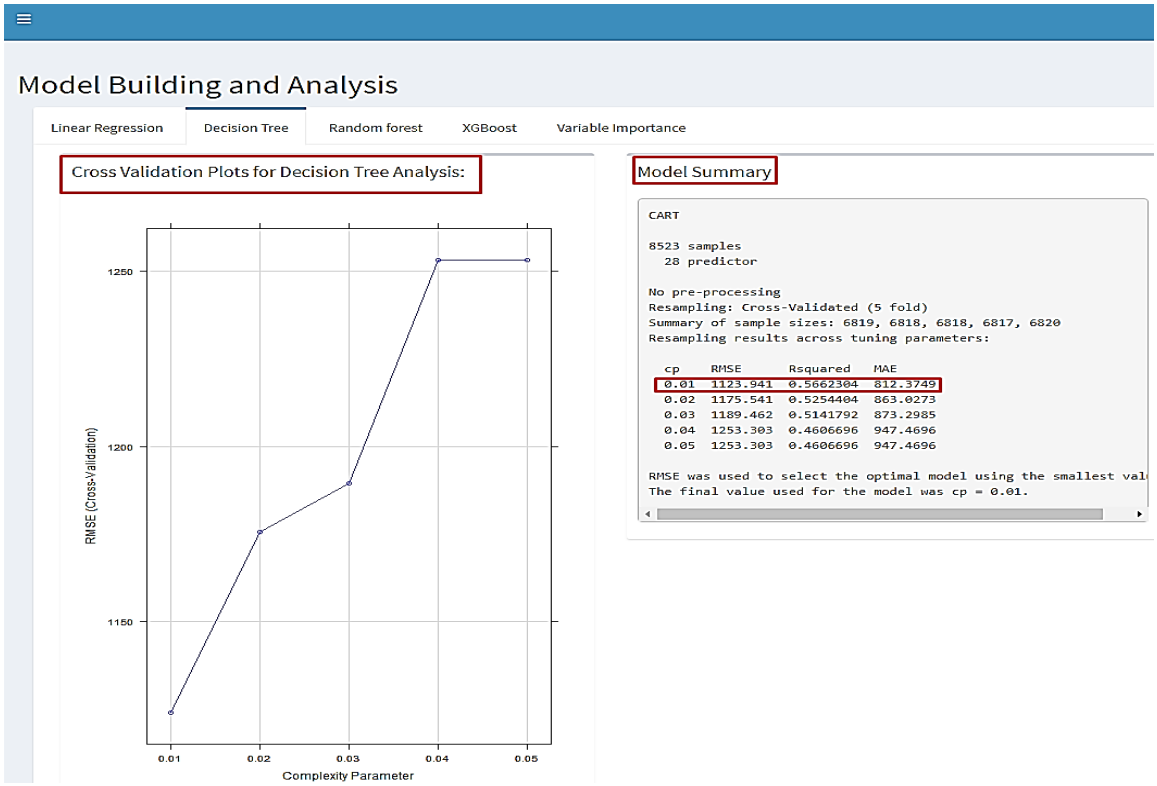


Fig 6.8. Decision Tree Model Analysis

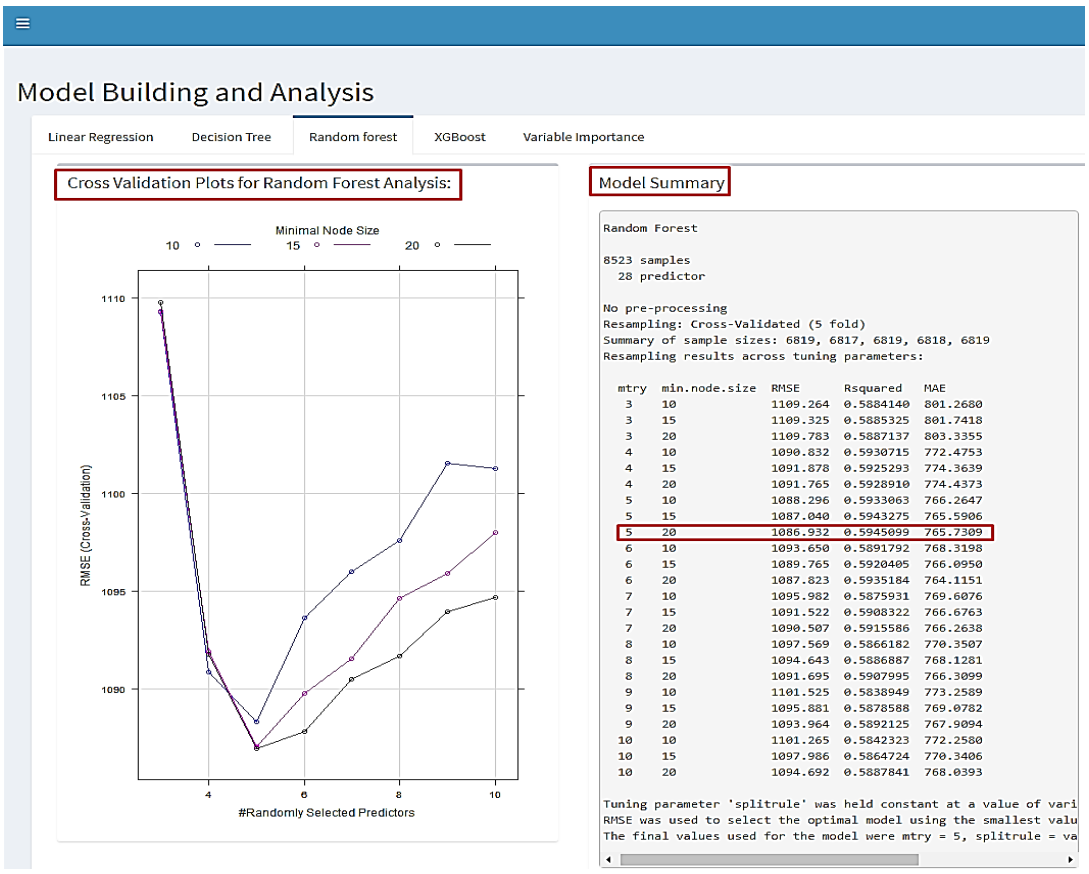


Fig 6.9. Random Forest Model Analysis

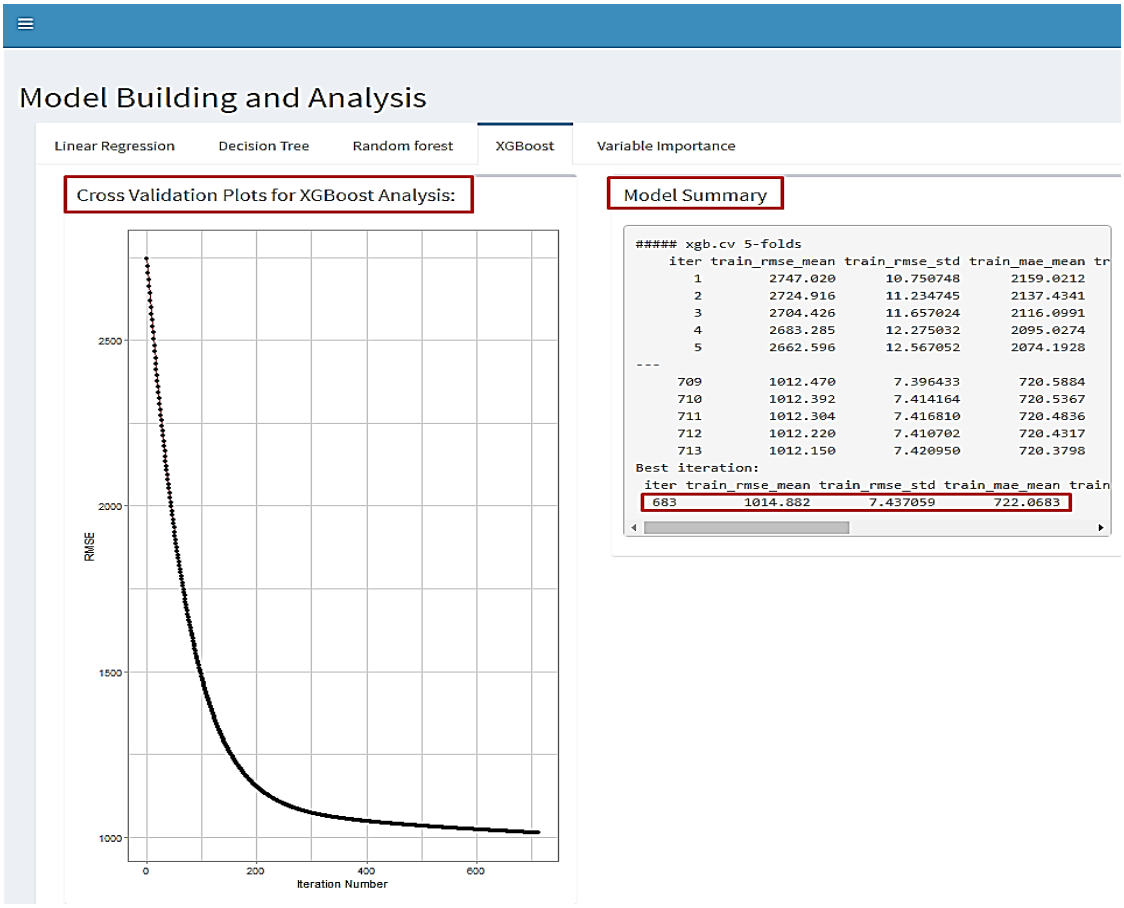


Fig 6.10: XGBoost Model Analysis

This Model Analysis interface also contains the variable importance tab showing in figure 6.11 which represents the important characteristics having a significant impact on sales forecasting.

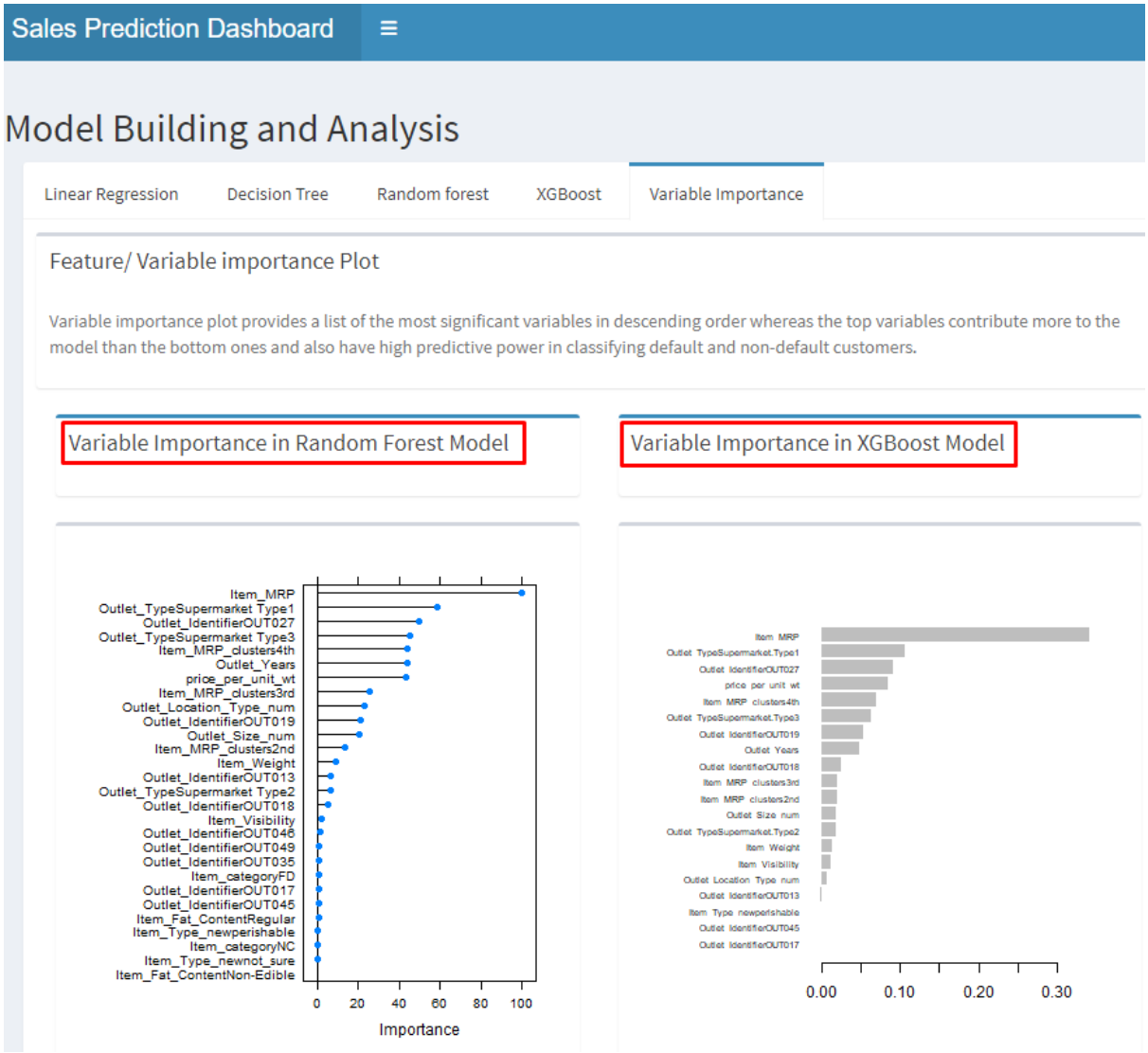


Fig 6.11: Dashboard view for Variable Importance

6.2.7 MenuItem-7: Prediction

As XGBoost gives the higher accuracy for the sales prediction model, this model has been implemented for the prediction use cases. In the sales prediction dashboard, if anyone want to preview the Item_Outlet_Sales of any Item_Identifier in a Specific Outlet_Identifier then the result will be displayed in accordance with the output of XGBoost Model on test Dataset as shown in figure 6.12 and 6.13.

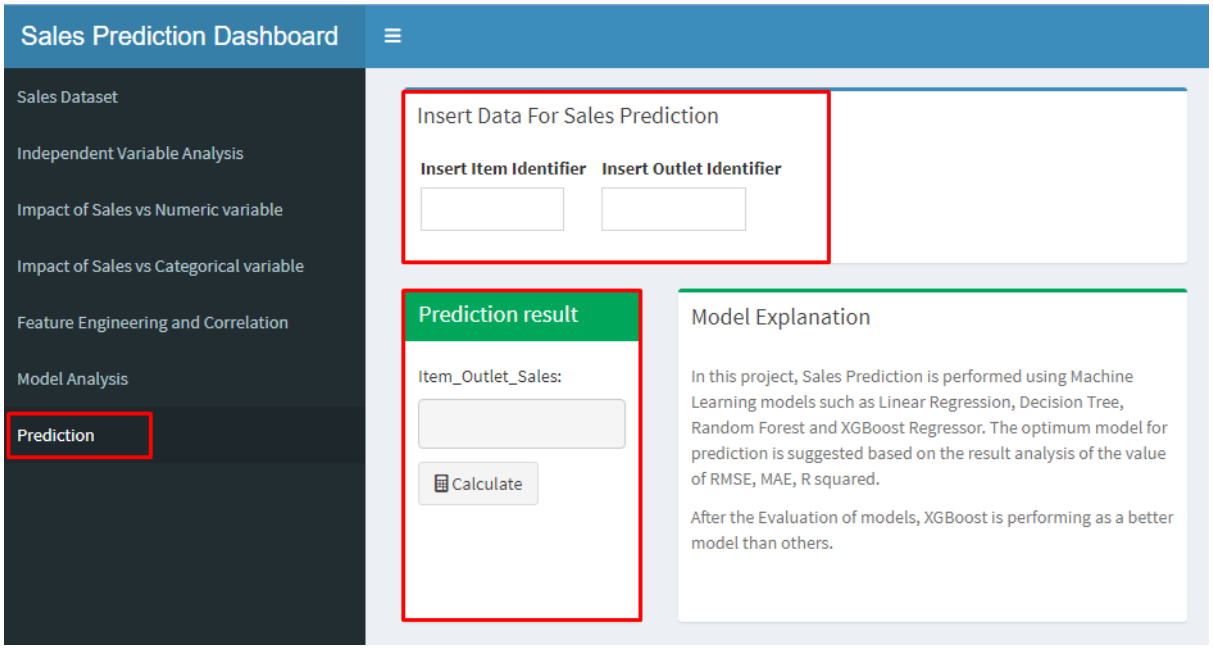


Fig 6.12: Dashboard view for Prediction

If user want to calculate one of the Item_Outlet_Sales, user can insert any Item_Identifier with the Outlet_Identifier then the result will be displayed after calculating with the best prediction model as shown in figure 5.24.

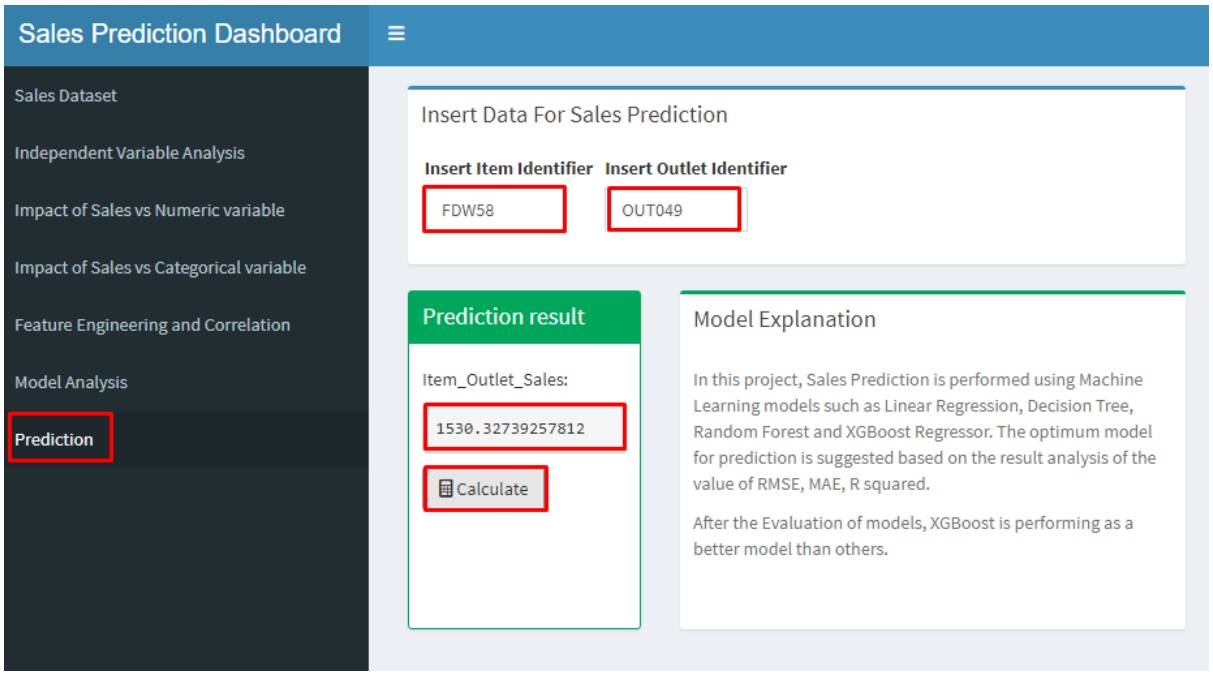


Fig 6.13: Dashboard view for Prediction result

If the input of Item_Identifier or Outlet_Identifier is wrong, then the result will be displayed as NA showing in figure 6.14.

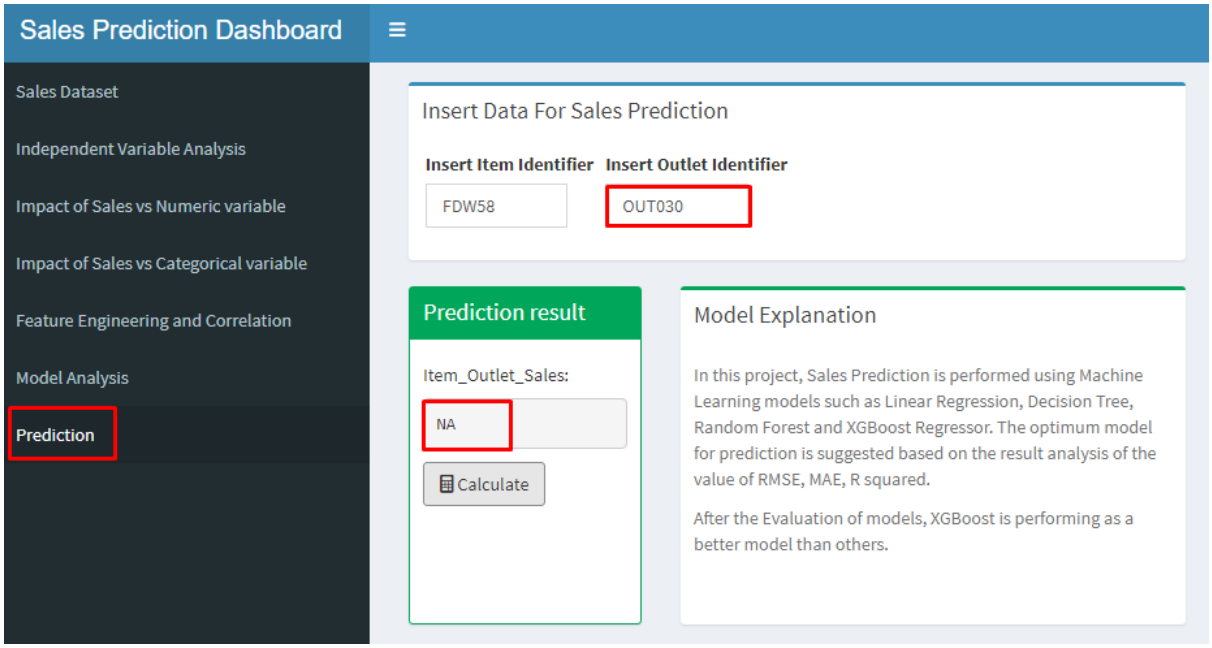


Fig 6.14: Dashboard view for Prediction result

CHAPTER 7
CONCLUSION

7.1 Conclusion

One of the more typical difficulties imposed on by current industry trends is dealing with a large amount of data. In light of the fact that data has grown to be one of the most important resources, SC managers are keen to gather key information that will provide them a strategic advantage. Utilizing machine learning techniques, users can examine vast volumes of data and find relationships between different elements that can lead to explicit knowledge and enhance the decision-making abilities of the user. In terms of accuracy or prediction, ML is a better tool from a business standpoint than conventional forecasting methodologies. It makes it possible to find hidden patterns that can be utilized as a starting point to find emerging market trends.

To avoid running out of sale items during any season, every shopping center in the modern world wants to know the customer demands in advance. Companies or shopping centers are getting better at anticipating the demand for goods sales day by day. For precise sales forecasting, extensive research is being conducted at the organization level. Because precise sales forecasts strongly correlate with a company's earnings. At the end of this study, it can say that it was primarily an interdisciplinary investigation where ML was applied to supply chain issues like sales forecasting. For the best level of sales prediction accuracy throughout this project, a number

7.2 Project Contributions

A more informed decision-making process in the related industry may be aided by the anticipated outcome of outlet item sales. It would assist management in making choices regarding how to use products and channels to improve the business. The application prototype's ability to help any professional in the industry visualize the dataset at a higher level was one of this project's significant achievements. This project can serve as a starting point for future business model in outlet item sales forecast in large-scale markets using data mining and ML algorithms.

7.3 Limitations of the Project

Though some important goals have been achieved in this thesis, there are still some limitations:

(a) The technologies utilized in this project, particularly the computer's processing power, were the main limitations of research restrictions. besides that, the research's time constraints.

(b) Ultimate results of the ML models highly depend on the prepared dataset. But to prepare an appropriate dataset, a lot of real time industry data sheets and survey is required. Once the appropriate dataset is prepared, only then maximum accuracy can be expected based on ML techniques. Our proposed ML models can act as a prototype solution for the industry practice.

(c) In order to create more accurate predictive models in the future, incorporating more explanatory variables can be used into the data.

(d) In addition, as stated in the objectives, our aim was to work with real-time dataset. But during this implementation work, considering the nature of sensitivity of the dataset, no company was willing to share the data feeds. As a result, we could not achieve the exact expected analytical results for the performance evaluation. Hence the lack of real-time data currently prevents a full analysis of the pipeline in the event of real-time data import. Therefore, a future study that uses both real-world data and simulation methodologies should look at the gathering, processing, and modeling of data in real-time.

(e) The implementation and integration of such an application into the business's sales and operations will provide a new set of difficulties that need also be researched in the future.

(f) This kind of research necessarily requires a certain level of experience in the relevant fields.

7.4 Future Works

There are numerous implications for being able to effectively estimate outlet item sales in this area of study utilizing advanced analytics. The forecast might contribute to a more informed decision-making process inside the organization. It would support management in their choice

to optimize products and outlet location in order to deliver a better customer experience, which would then result in an increase in sales, which would most likely result in an increase in revenue and cause the organization to expand. This project's contribution to a greater understanding of the algorithms and data mining techniques is one of its most significant accomplishments.

For further research this project can be re-run with different sets of data using different ML models in terms of reduction of computational cost. Furthermore, this application can expand with the more efficient ML algorithms to integrate with the industry.

REFERENCES

Bajaj, P., Ray, R., Shedge, S., Vidhate, S., and Shardoor, N. k. (2020). Sales Prediction using Machine Learning Algorithms, International Research Journal of Engineering and Technology (IRJET), vol. 07, no. 6, pp. 3619-3625.

Behera, G., and Nain, N. (2019). A comparative study of big mart sales prediction. International Conference on Computer Vision and Image Processing, Springer, Singapore, pp. 421-432.

Cadavid, J. P. U., Lamouri, S., and Grabot, B. (2018). Trends in Machine Learning Applied to Demand & Sales Forecasting: A Review, International Conference on Information Systems, Logistics and Supply Chain.

Chandel, A., Dubey, A., Dhawale, S., and Ghuge, M. (2019). Sales Prediction System using Machine Learning, International Journal of Scientific Research and Engineering Development (IJSRED), vol. 2, no. 2, pp. 667-670.

Cheriyan, S., Ibrahim, S., Mohanan, S., and Treesa, S. (2018). Intelligent Sales Prediction Using Machine Learning Techniques, International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 53-58.

Dairu, X., and Shilong, Z. (2021). Machine Learning Model for Sales Forecasting by Using XGBoost, International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 480-483.

Decision Tree Classification Algorithm. [Online]. Available: https://www.javatpoint.com/ machine-learning-decision-tree-classification-algorithm

Gurnani, M., Korke, Y., Shah, P., Udmale, S., Sambhe, V., and Bhirud, S. (2017). Forecasting of sales by using fusion of machine learning techniques, International Conference on Data Management, Analytics and Innovation (ICDMAI), pp. 93-101.

Iakovou, S. A., Kanavos, A., and Tsakalidis, A. (2016). Customer Behavior Analysis for Recommendation of Supermarket Ware, 12th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), Thessaloniki, Greece, pp. 471–480, 16–18.

Kadam, H., Shevade, R., Ketkar, D., and Rajguru, S. (2018). A forecast for big mart sales based on random forests and multiple linear regression, Int. J. Eng. Dev. Res., vol. 6, no. 4, pp. 41–42.

Kumar, Saravana, NM., Hariprasath, K., Kaviyavarshini, N., and Kavinya, K. (2020). A Study on the Forecasting Bigmart Sales Using Optimized Data Mining Techniques, Science in Information Technology Letters 1, no. 2, pp. 52-59.

Linear Regression with Sales Prediction Project. [Online]. Available: https:// medium.com/-@jagwithyou/linear-regression-with-sales-prediction-project-8152e7de2cf2

N, M., Chatradi, P., V, A. C., Kalavala, S. M., and S, N. K. (2020). Improvizing big market sales prediction, J. Xi'an Univ. Archit. Technol., vol. XII, no. IV, pp. 4307–4313.

Narkhede, A., Awari, M., Gawali, S., and Mhaisgawali, A. (2020). Big mart sales prediction using machine learning techniques, Int. J. Sci. Res. Eng. Dev., vol. 3, no. 4, pp. 693-697, 2020.

NIYA, N. J., and Jasmine, J. (2021). Sale Prediction using linear regression model, International Journal of Creative Research Thoughts (IJCRT), vol. 9.

Punam, K., Pamula, R., and Jain, P. K. (2018). A two-level statistical model for big mart sales prediction, International Conference on Computing, Power and Communication Technologies (GUCON), pp. 617–620.

Sadia, K. H., Sharma, A., Paul, A., Padhi, S., and Sanyal, S. (2019). Stock market prediction using machine learning algorithms, Int. J. Eng. Adv. Technol., vol. 8, no. 4, pp. 25–31.

Sengar, R. S., and Ahmed, D. S. F. (2019). Review on Trends in Machine Learning Applied to Demand & Sales Forecasting, IJOSCIENCE, vol. 5, no. 12, pp. 25–29.

Seyedan, M., and Mafakheri, F. (2020). Predictive Big Data Analytics for Supply Chain Demand Forecasting: Methods, Applications, and Research Opportunities, Journal of Big Data, vol. 7, no. 53, pp. 1-22.

Shah, D., Isah, H., and Zulkernine, F. (2019). Stock market analysis: a review and taxonomy of prediction techniques, Int. J. Financ. Stud., vol. 7, no. 2, p. 26.

Shinde, P. P., Oza, K. S., and Kamat, R. K. (2017). Big data predictive analysis: Using R analytical tool, International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 839–842.

Theresa, I., Medikonda, V. R., and Reddy, N. (2020). Prediction of Big Mart Sales Using Exploratory Machine Learning Techniques, International Journal of Advanced Science and Technology, vol. 29, no. 6, pp. 2906-2911.

XGBoost. [Online]. Available: https://www.geeksforgeeks.org/xgboost/