Bachelor of Science in Computer Science
May 2020

Predicting sales using Machine Learning Techniques

Sai Nikhil Boyapati
Ramesh Mummidi

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Sai Nikhil Boyapati
E-mail: sabo19@student.bth.se

Ramesh Mummidi
E-mail: ramu19@student.bth.se

University advisor:
Suejb Memeti
Department of computer science (DIDA)

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

Abstract

Background: Sales forecasting is an important field in the food sector, and it has recently got immense popularity to boost market operations and productivity due to new technologies. The industry has traditionally focused on a conventional statistical model but in the recent years, Machine Learning techniques have received more attention.
Objectives: This thesis will help to identify the critical features that influence sales and also an experiment is performed to find the best suitable algorithm for sales forecasting.
Methods: Machine Learning Algorithms such as Simple Linear Regression, Gradient Boosting Regression, Support Vector Regression, and Random Forest Regression were considered in this thesis, which they expected to perform well on the issues. An experiment is carried out to determine the efficiency of the algorithms.
Results: Algorithms such as Simple Linear Regression, Gradient Boosting Regression, Support Vector Regression, and Random Forest Regression are commonly known for performing better than others, this has been clearly shown that Random Forest Regression is the most appropriate algorithm compared to the others.
Conclusions: The Random Forest Regression algorithm performed well after doing all the study when compared with other algorithms. Hence the Random Forest Regression is considered as the best suitable algorithm for forecasting product sales.

Keywords: Correlation, Machine Learning, Performance Metrics, Sales Forecasting

Acknowledgments

First and foremost, praises and thanks to God, the Almighty, for His showers of blessings throughout our research work to complete the research successfully. We would like to express our deep and sincere gratitude to our supervisor Suejb Memeti, for giving us the opportunity to do research and providing invaluable guidance throughout this research. Under his guidance it was a great privilege and honour to work and study. We are extremely appreciative of what he has offered us.

Finally, our special thanks to all the people who have helped us directly or indirectly in completing the research work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Earlier companies used to produce goods without considering the number of sales and demand. For any manufacturer to determine whether to increase or decrease the production of several units, data regarding the demand for products on the market is required. Companies can face losses if they fail to consider these values while competing on the market. Different companies choose specific criteria to determine their demand and sales [1].

In today's highly competitive environment and ever-changing consumer landscape, accurate and timely forecasting of future revenue, also known as revenue forecasting, or sales forecasting, can offer valuable insight to companies engaged in the manufacture, distribution or retail of goods[2]. Short-term forecasts primarily help with production planning and stock management, while long-term forecasts can deal with business growth and decision-making[1].

Sales forecasting is particularly important in the industries because of the limited shelf-life of many of the goods, which leads to a loss of income in both shortage and surplus situations. Too many orders lead to a shortage of products and still too few orders lead to a lack of opportunity. Therefore, competition in the food market is continuously fluctuating due to factors such as pricing, advertisement, increasing demand from the customers[3].

Managers usually make sales predictions randomly. Professional managers, however, become hard to find and not always available (e.g., they can get sick or leave). Sales predictions can be assisted by computer systems that can play the qualified managers' role when they are not available or allow them to make the right decision by providing potential sales predictions. One way of implementing such a method is to try and model the professional managers' skills inside a computer program[4].

Alternatively, the abundance of sales data and related information can be used through Machine Learning techniques to automatically develop accurate sales predictive models. This approach is much simpler. It is not prejudiced by a single sales manager's particularities and is flexible, which means it can adapt to data changes. It has, however, the potential to overestimate the accuracy of the prediction of a human expert, which is normally incomplete. For example, once companies used to produce the products without taking into consideration the number of sales and demand as they faced several problems. Since they don't know how much to sell, for any manufacturer to decide whether to increase or decrease the number of units, data regarding the consumer demand for products is essential. If companies do not consider these principles when competing in the market, they will face losses. Different companies choose different parameters to determine their market and sales.

There are several ways of forecasting sales in which companies have previously focused on various statistical models such as time series and linear regression, feature engineering and random forest models to obtain future sales and demand prediction. Time series contains data points that are stored over a fixed period and are used to forecast the future. Time series is a collection of data points which are collected in period at sequential, evenly spaced points. The most important components to analyze are patterns, seasonality, irregularity, cyclicity.

Linear regression is a mathematical tool used to forecast past values. It can help to determine the underlying trends and address cases involving overstated rates[5][6]. Feature engineering is the use of data on domain knowledge and the development of features to make predictive Machine Learning models more accurate. It makes for deeper data analysis and a more useful perspective[7]. A decision tree is a fundamental principle behind a model of random forests. The decision tree approach is a technique used in data mining to forecast and classify data. The decision tree approach does not provide any conceptual understanding of the issue itself. Random forest is the more sophisticated method that allows and merges many trees to make decisions. The random forest model results in more accurate forecasts by taking out an average of all individual tree decision predictions.

The entire data set is usually divided into two parts, namely the training data and the test data. Training data is a data that is used to train the model, and test data is the data used to evaluate the trained model. A classical approach is 80-20 split, stating that 80 percent of the data is used to train the model, and the remaining 20 percent of the data is used to test the model. But approaches like stratified K-fold cross-validation are known to provide good results. There were many cross-validation variants, such as simple k-folds, leave one out, stratified k-fold cross-validation, and so on[8][9].

1.0.1 Aims and Objectives

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

Objectives:

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

1.0.2 Research Questions

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

RQ1:

What are the critical features that influence product sales?

Motivation:

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

RQ2:

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

Motivation:

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

1.1 Background

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

1.1.1 Data Mining

Data mining is described as a process for extracting usable data from a larger collection of raw data using statistical, artificial intelligence, Machine Learning and pattern recognition methods[10][11]. Data Mining is increasingly seen as a step in a systematic and iterative process of knowledge discovery, in which automated pattern recognition methods are combined with expert knowledge of the analyst. This process is called the Knowledge Discovery in Databases (KDD) process[12].

1.1.2 Machine Learning

Machine Learning is the area of study which enables machines to learn without being explicitly programmed[13]. Machine Learning is defined as the computer program learns from experience E with respect to some class of tasks T and performance measure P when its performance at tasks in T, as measured by P, strengthens with

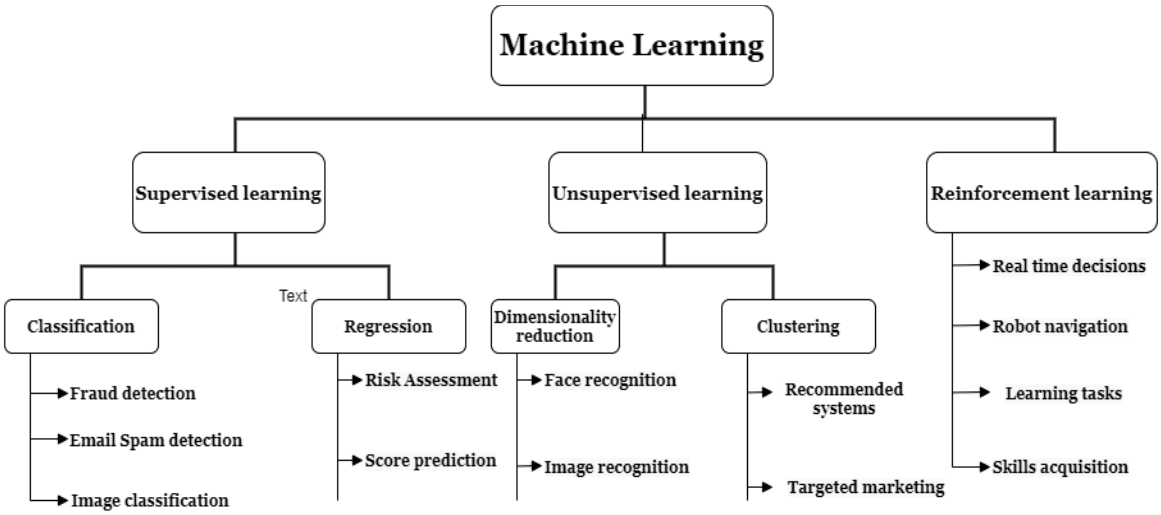

Figure 1.1: Types of Machine Learning

experience E[14]. In general, Machine Learning is a program that can manage various tasks by analyzing and exploring data[15].

Common Machine Learning applications such as email spam detection, credit card fraud, stock predictions, smart assistants, product recommendations, self-driving cars, sentiment analysis, etc.

Supervised Learning:

The most popular model for performing Machine Learning processes is supervised learning. It is commonly used for data where the mapping between input-output data is accurate. Supervised learning is the subset of Machine Learning which concentrates on learning a model of classification or regression, that is, learning from labeled test data[15].

Unsupervised Learning:

The data is not explicitly labeled into different classes in the case of unsupervised learning that is there is only unlabeled data. By identifying implicit patterns the model can learn from the data. Unsupervised Learning categorizes the densities, structures, related segments, and other similar properties based on the data[16].

Reinforcement Learning:

Reinforcement Learning is a sub-field of Machine Learning. In a given scenario, it is about taking appropriate action to optimize reward. Various algorithms and computers are employed to determine the best possible action or path it will follow in a specific scenario. Reinforcement learning varies with supervised learning in such a way that the training data has the answer key with it in supervised learning such that the model is trained with the correct response itself while in reinforcement learning there will be no response but the reinforcement agent determines how to

execute the task. It is required to learn from its experience, in the absence of training data[15].

1.1.3 Machine Learning Algorithms:

Forecasting means predicting events of the future, typically based on previous records. For a long time, statistical models were commonly used for the conducting of predictions. The role of generalization in Machine Learning has been considered. In the case when a new product or store is introduced, this effect could be used to make sales predictions because there is a limited amount of historical data for a particular time series[17]. In this thesis we have used supervised learning algorithms such as Support Vector Machines, Random Forest Regression, Gradient Boosting, and Simple Linear Regression. These can make it easier to find better outcomes compared to traditional analytical techniques of time series[18].

Simple Linear Regression

Simple linear regression is useful for defining a relationship between two continuous variables. One is an indicator or independent variable and another is an answer or dependent variable. It looks for a statistical relationship, but not a deterministic one. The relationship between the two variables is said to be deterministic if one variable can be precisely represented by the other[19]. For example, it is possible to correctly forecast Fahrenheit by using temperature in degree Celsius. The mathematical equation is not sufficient to assess the association between the two variables. For example, the relationship between weight and height. The Equation for the Simple linear regression is:

$$Y = a + bX$$

where Y is the expected value of the dependent variable y for every specified value of the independent variable X, a is the intercept, b is the regression coefficient and X is independent variable.

Gradient boosting Regression

Gradient boosting is some kind of enhancement in Machine Learning. It is based on the premise that, when combined with previous ones, the best possible current iteration will minimize the maximum prediction error. The key idea for this next iteration is to set the target outcomes to minimize the error.

One of the most successful Machine Learning models for predictive analytics is the Gradient Boosted Regression Trees (GBRT) model (also called Gradient Boosted Machine or GBM), which makes it an industrial workhorse for Machine Learning. The Boosted Trees Model is a type of additive model that combines decisions from a sequence of base models to make predictions[20]. One can write this class of models more formally, as:

$$g(x) = f0(x) + f1(x) + f2(x) + f3(x) + ...$$

Where the final classifier g is the amount of the specific classifiers fi. Each base classifier is a simple decision tree for model boosted trees. This broad approach of using multiple models is called model ensembling to achieve better predictive performance. Unlike Random Forest, which independently builds all the base classifiers, each using a subsample of data, GBRT uses a particular technique of assembly called gradient boosting[20].

Support Vector Machine

Support Vector Machine or SVM is one of the most common Supervised Learning algorithms used for both Classification and Regression issues. The SVM algorithm aims to build the best line or decision boundary that can divide n-dimensional space into conveniently place the new data point in the right category in the future. The optimal choice boundary is called a hyper plane. SVM chooses extreme points vectors that help to create a hyper plane. Such extreme cases are called help vectors[21]. The equation for Support Vector Regression is:

$$f(x) = x'\beta + b$$

Random Forest Regression

Random Forest is one of the most powerful Machine Learning frameworks for predictive analytics. A random forest method is a type of discrete structure that allows predictions by integrating decisions from a series of simple models[22]. More formally, this subset of models can be written as:

$$g(x) = f0(x) + f1(x) + f2(x) + f3(x) + ...$$

Where the initial configuration g is the number of the initial specific model's fi. Here, any base classifier is a simple decision tree. This wide-ranging technique of using multiple models to improve predictive performance is called model assembling. In random woods, all baseline models are built independently using a separate subset of results.

1.1.4 Selection of Machine Learning Algorithms

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

1.1.5 Selection of Performance Metrics

To identify the appropriate algorithm one needs to evaluate the results and then we can predict it. In this case, the accuracy score would play a crucial role while measuring the performance of the algorithm. For calculating the average magnitude of errors mean absolute error metric will be used in this study. In real-time data there might be a chance of worst-case error between the actual value and predicted value in this particular scenario max error is used.

Chapter 2

Related Work

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

Due to the importance of forecasting in various fields, there are so many different types of approaches taken previously, some of the methods such as Machine Learning models, hybrid models, and statistical models. To handle this work, some of the statistical methods such as auto regressive moving average (ARMA) and auto regressive integrated moving average (ARIMA) will be helpful[4].

İrem İşlek and Şule Gündüz Öğüdücü experimented with the use of bipartisan graphic clusters that clustered different warehouses according to the sales behavior. They addressed the application by applying the Bayesian network algorithm in which they managed to produce the enhanced forecasting experience[23].

Grigorios Tsoumakas had used Machine Learning techniques to perform a survey on the forecasting of food sales. They had addressed data analyst design decisions such as temporal granularity, output variable, and input variables in this survey[4]. In this paper the authors experimented by taking the point of sale (POS) as internal data and even external data by considering different environments to enhance the efficiency of demand forecasting. They considered different Machine Learning algorithms such as Boosted Decision Tree Regression, Bayesian Linear Regression, and Decision Forest Regression for evaluation[24].

The paper's authors had researched interestingly about customers coming to the restaurants using Random Forests, k-nearest neighbor, and XGBoost. They chose two real-world data sets from different booking sites and also made different input variables from restaurant features. The results have shown that XGBoost is the most appropriate model for the dataset[25].

Holmberg and Halldén had observed that regular restaurant sales to be influenced by the weather. They considered two Machine Learning algorithms as XGBoost and neural network, and the results showed that the XGBoost algorithm is more accurate than the other algorithm, and they also found that they had improved their model performance by 2-4 percentage points by taking weather factors into consideration. To improve accuracy, they had considered numerous variables such as date characteristics, sales history, and weather factors[26].

Most of the recent studies focused on sales modeling without considering the relationship between the training and testing data, they used training data directly. This causes many errors which lead to a reduction in accuracy. Recent studies have suggested clustering techniques to separate the entire forecasting data into

several clusters of predictable data before designing predictable models to minimize computational time and achieve effective evaluating performance[27].

In particular, Support Vector Machine(SVM) had been applied to demand forecasting. Garcia et al. (2012), in their study, proposed an intelligent model that relies on supporting vector machines to deal with issues relating to the allocation and revelation of new models. Kandananond (2012) showed that SVM surpassed Artificial Neural Networks in estimating demand for consumer goods[28].

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

Chapter 3

Method

In this thesis research questions are answered by using research methods. Research aspects for this work are examined by the execution of the experiments.

3.1 Experiment

An experiment is chosen for the first research question i.e. correlation. Each data attribute can be selected by applying feature selection methods like data correlation and which will make the predictable attributes more accurate. This will reduce a lot of strain on the Machine Learning model during pre-processing and cleansing the data. For the second research question an experiment is chosen because the experiments provide control over factors and a deeper understanding of many common research techniques such as a case study or survey[29]. One can describe the procedure followed in this experiment as follows:

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

3.2 Experimentation Environment

Python

Python is a commonly used high-level programming language, it was designed by Guido van Rossum which can be easy to interpret and read[30]. Python has specific functionality and is convenient to be used for both quantitative and analytical computational purposes. Data Science Python is popularly used and, as well as being a dynamic and open source language, is a top choice. Its massive libraries are also used to manipulate the data however for a beginner data analyst they are really simple to learn[31]. The python libraries used in this thesis are briefly described as follows:

NumPy

NumPy is a library that consists of multidimensional array objects and a set of array processing routines. NumPy is used along with SciPy and Matplotlib packages. This combination is used for technical computing. Mathematical and logical operations are performed with the help of NumPy[32].

Pandas

Pandas is a software library that is designed for manipulating the data and analysis in a python programming language. It is open-source which is released under the BSD license of three clauses. It is based on the Numpy package, and the DataFrame is its main data structure[33].

Matplotlib

Matplotlib is a module of Python used to plot the attractive Graphs. Visual representation in data science is a significant step. One can quickly understand how data is split by using visual representation. There are many libraries to represent the data, but the matplotlib is very widely known and easier to visualize[34].

SKlearn

Scikit-learn is a free python library. It features multiple clustering classification and regression algorithms including random forests, DBSCAN, k-means, gradient boosting, support vector machines, and gradient boosting which is programmed to interface with the NumPy and SciPy libraries[35].

Seaborn

Seaborn is a open-source python library that is used for statistical graphics. It offers a data set-oriented API to analyze relationships among different variables, as well as resources to select color palettes that truly in the data[36].

3.3 Data overview

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

3.4 Feature Selection

There are various types of factors that can make the model of Machine Learning more effective on any given task. One of the methods of feature selection is data correlation

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

Figure 3.1: Dataset description

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

Figure 3.2: Dataset summary

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

Figure 3.3: Correlation values

which will have a major impact on the model's performance. This will reduce a lot of strain on the Machine Learning model during preprocessing and cleansing the data. The data attributes chosen for training the Machine Learning model would have a major impact on the efficiency of the model. Because of the irrelevant features that are presented, the model output will be reduced. The feature selection method provides an efficient way to remove data redundancy and irrelevant data that helps to reduce computation time, improve accuracy, and also enhance understanding of the model[37].

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

3.4.1 Data Correlation Method

Data correlation is a method that helps to predict one attribute from another attribute and is used as a basic quantity in many modeling techniques. If one feature increases, the correlation will be positive, so the other feature increases as well and negative if one feature increases there will be a reduction in another. If there is no relation between any two attributes then it is said to be no correlation[38]. If there is a linear relationship between the constant variables then the Pearson correlation coefficient is used. If there is a non-linear relation between the constant variables then the Spearman correlation coefficient is used.

Since the considered data set is linear so the Pearson correlation coefficient is used for the selection of features in this study. This correlation for all the attributes is shown in figure 3.4. To improve the efficiency of the Machine Learning model, the attributes that have negative correlations were removed. It is a statistic measuring the linear correlation of two variables X and Y. It has a value between +1 and 1, where 1 is a linear positive correlation, 0 is not a linear correlation and 1 is a linear negative correlation[39].

The motivation for considering the correlation is when people know a score on one measure, they can make a prediction of another measure that is highly related to it more accurate. The more accurate the prediction, the stronger the relationship between the variables.

The heat map for correlation between non-numerical attributes is plotted as follows:

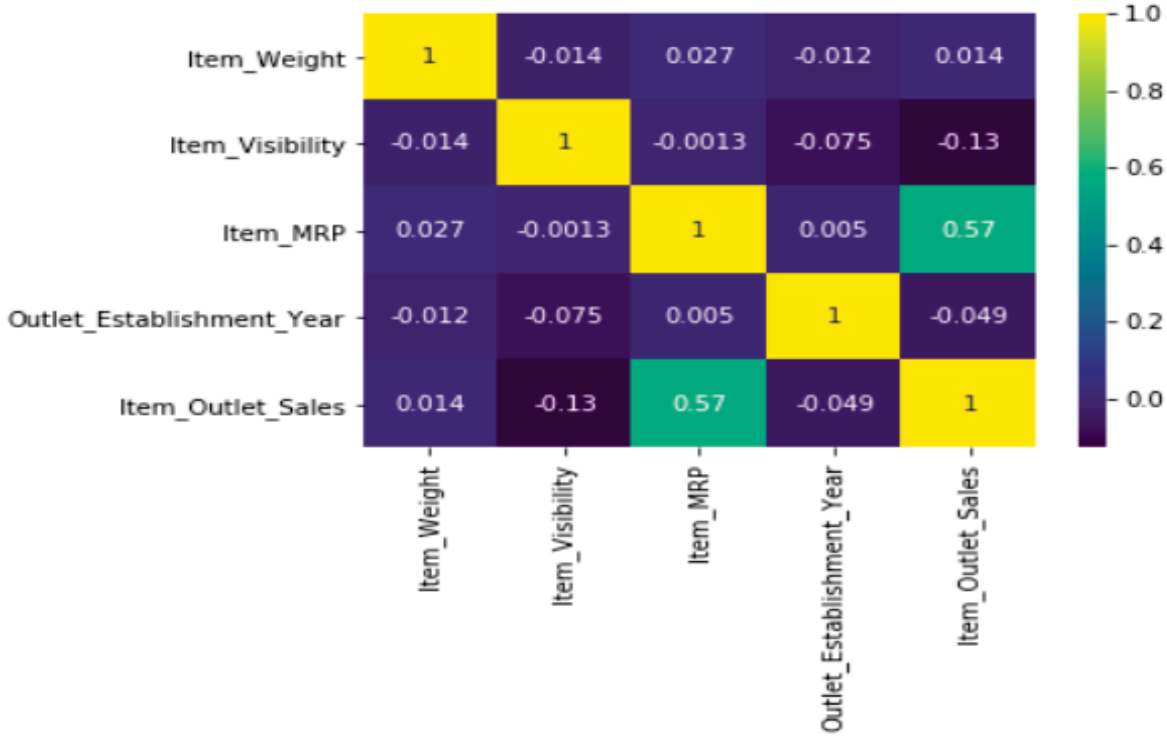

Figure 3.4: Heat map

3.5 Feature Importance

Feature Importance refers to a class of approaches for assigning values to input features to a predictive model which determines the relative significance of each factor while forecasting[40].

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

Feature Importance can be used to enhance a predictive model. This can be accomplished by selecting those features to remove (lowest scores) or those features to retain, using the importance scores. This is a type of selection of features, and can simplify the modeling problem, accelerate the modeling process, and in certain cases improve model performance[40].

3.6 Data preprocessing

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

3.6.1 Encoding Categorical Values

Categorical data contains label values that are considered nominal values. Each value has categories of different types. Besides, a few of the groups have a normal relationship with each other is known as natural ordering. The categorical data can be converted into numerical data to improve the efficiency of the Machine Learning model[41].

One Hot Encoding

One hot encoding is the method where the data is represented in binary format and included as a feature. It is one of the most common methods, comparing each level of the numerical variable with a fixed starting point. In this thesis for the data set that had taken, one hot encoding is used to represent categorical variables as binary vectors[41].

This approach results in a dummy variables trap because it is easy to predict the outcome of one variable with the support of the existing variables. This trap leads

h

Outlet_Size	Outlet_Location_Type
Medium	Tier 1
Medium	Tier 3
Medium	Tier 1
Medium	Tier 3
High	Tier 3
...	...
High	Tier 3
Medium	Tier 2
Small	Tier 2
Medium	Tier 3
Small	Tier 1

Figure 3.5: Before One Hot Encoding

Outlet_Size_High	Outlet_Size_Medium	Outlet_Size_Small
0	1	0
0	1	0
0	1	0
0	1	0
1	0	0
...	...	...
1	0	0
0	1	0
0	0	1

Figure 3.6: After One Hot Encoding

to a multicollinearity problem. It occurs when the independent features become dependent upon each other. To overcome the multicollinearity problem one of the dummy variables should be dropped. The following figure represents the before and after one hot coding.

3.6.2 Stratified K-fold Cross-Validation

Cross-validation (CV) is a procedure of statistical analysis used to assess the effectiveness of a Machine Learning technique, as well as a re-sampling method used to validate an algorithm if there is insufficient data[42].

Stratification is the process of rearranging the data to ensure each fold is a good representative of the whole. Data splitting into folds may be controlled by criteria such as ensuring where each fold has the same ratio of outcomes with a given categorical value, such as the class outcome value. This process is called stratified k-fold cross-validation[43]. Common techniques of cross-validation include K-fold cross-validation, Stratified K-fold cross-validation, and cross-validation leave-one-out.

The motivation behind the 10-fold stratified cross-validation is that the estimator has a lower variance than a single hold-out set estimation method which could be very essential if there is a limited amount of data. There will be plenty of variance in the results estimate for various data samples, or for specific data partitions to create training and test sets. The 10-fold stratified cross-validation removes this variance by comparing more than 10 separate partitions, thereby making the performance estimate less sensitive to data partitioning.

3.7 Performance Metrics

Several metrics can be used while evaluating how well a model is performing. It is necessary to understand how each metric measures to select the evaluation metric to better assess the model. This thesis main objective was to compare the performance of Machine Learning techniques by evaluating all of these performance metrics such as Accuracy score, Mean Absolute Error, and Max error.

3.7.1 Accuracy score

Accuracy is known as the ratio of a several correct predictions(both true positives and true negatives) to the total number of data points[44].

$$AccuracyScore = \frac{FN + FP}{N}$$

Where FN is false negative, FP is false positive and N is total number of predictions.

3.7.2 Max Error

The function max error measures the maximum standard errors, a metric representing a worse-case error between the expected value and the actual value. Max error

would be 0 on the test set in a properly fitted single-output regression analysis, and while this would be extremely impossible in the modern world, this measurement indicates the amount of error the model has when it was placed in[45].

$$MaxError(y, x) = max(|y_i - x_i|)$$

Where Yi describes the actual values, Xi describes the expected values.

3.7.3 Mean Absolute Error

Mean Absolute Error is a process performance measure that is used for regression models. A model's mean absolute error concerning a test data set is the average of the actual values on all instances in the test set of the specific prediction errors. For instance, every predictive error is the difference between the predicted value and the actual value. Mean Absolute Error is one of several metrics for summing up and measuring the Machine Learning model's performance[46].

$$MAE(y, x) = (\frac{1}{n_{samples}}) \sum_{i=0}^{n_{samples}-1} |y_i - x_i|$$

Where Yi describes the actual values, Xi describes the expected values.

Chapter 4

Results

Simple Linear Regressor, Gradient Boosting Regressor, Random Forest Regressor and Support Vector Regressor are trained with the set of data using a 10-fold stratified cross-validation approach that dynamically selected the training and testing with fixed proportion each time and the efficiency was calculated using max error, mean absolute error and accuracy metrics.

4.1 Simple Linear Regressor

The box plot in Figure 4.1 shows the accuracy score(ACC) obtained by the Simple Linear Regressor during a 10-fold stratified cross-validation test. The upper box-plot represents a maximum accuracy of 84.099 percent, the middle quartile represents a median accuracy of 81.2 percent, and the lower quartile of the box-plot represents a minimum accuracy of 73.95 percent.

The box plot in Figure 4.1 shows the Mean Absolute Error(MAE) obtained by Simple Linear Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum MAE of 4.6773, the middle quartile represents a median MAE of 3.2127 and the lower quartile of the box plot represents a minimum MAE of 2.5208.

The box plot in Figure 4.1 shows the Max Error(ME) obtained by Simple Linear Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum ME of 0.5118, the middle quartile represents a median ME of 0.4917 and the lower quartile of the box plot represents a minimum ME of 0.4731.

4.2 Gradient Boosting Regressor

The box plot in Figure 4.4 shows the accuracy score(ACC) obtained by Gradient Boosting Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum accuracy score of 91.2 percent, the middle quartile represents a median accuracy score of 86.27 percent and the lower quartile of the box plot represents a minimum accuracy score of 78.3 percent.

The box plot in Figure 4.5 shows the Mean Absolute Error(MAE) obtained by Gradient Boosting Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum MAE of 4.43, the middle quartile represents a median MAE of 3.15 and the lower quartile of the box plot represents a minimum MAE of 2.77.

The outcomes that Simple Linear Regressor obtains are as follows:

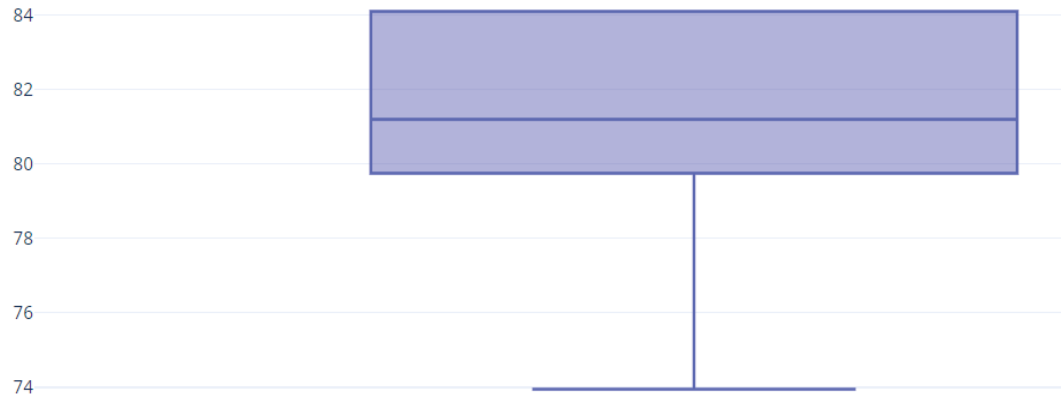

Figure 4.1: Accuracy Box plot for Simple Linear Regressor

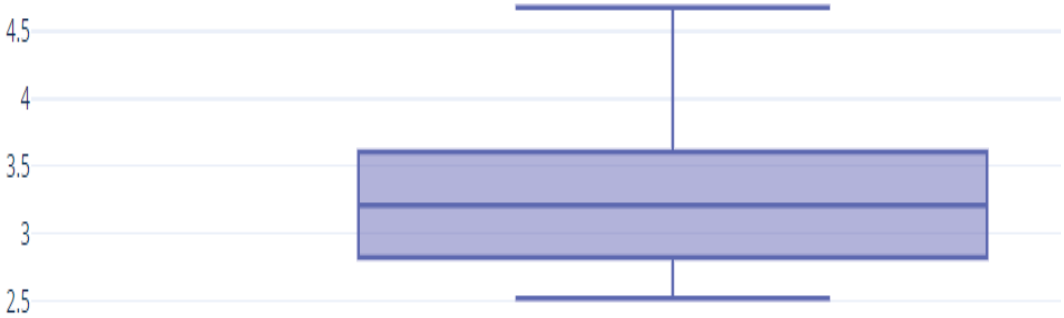

Figure 4.2: MAE Box plot for Simple Linear Regressor

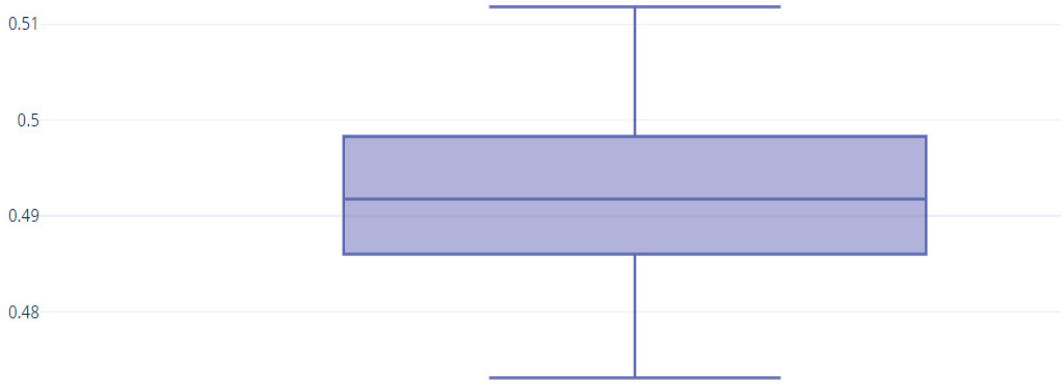

Figure 4.3: ME Box plot for Simple Linear Regressor

The outcomes that Gradient Boosting Regressor obtains are as follows:

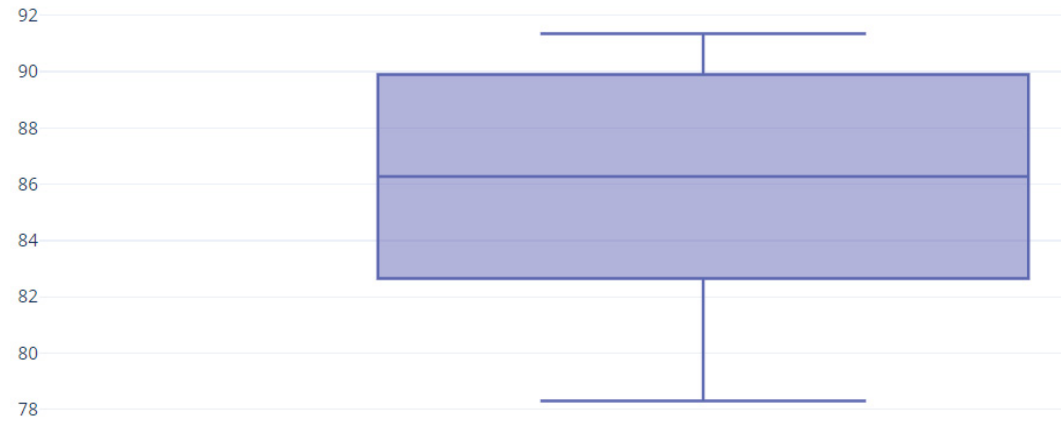

Figure 4.4: Accuracy Box plot for Gradient Boosting Regressor

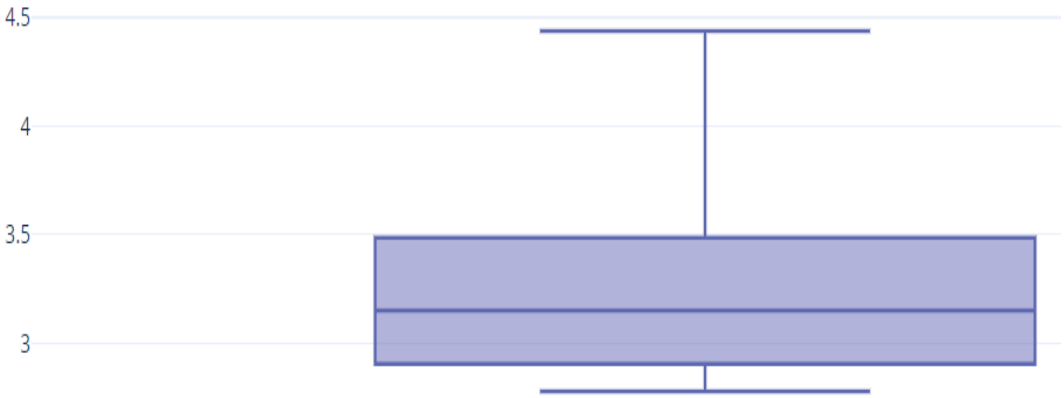

Figure 4.5: MAE Box plot Gradient Boosting Regressor

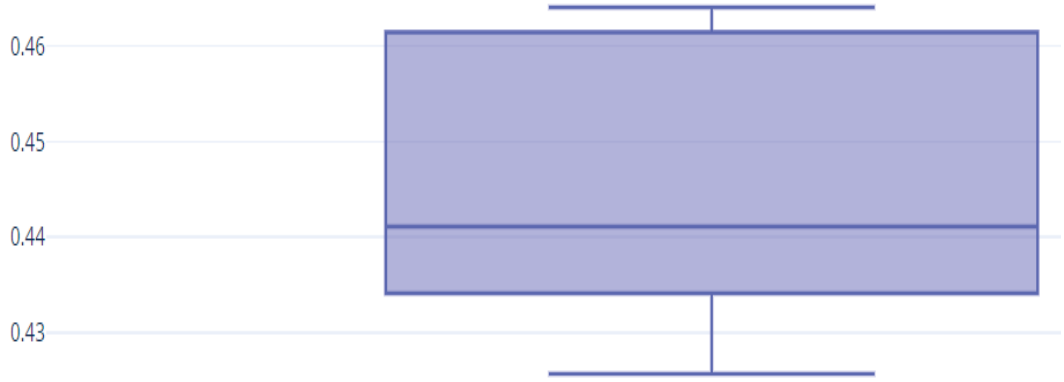

Figure 4.6: ME Box plot Gradient Boosting Regressor

The outcomes that Support Vector Regressor obtains are as follows:

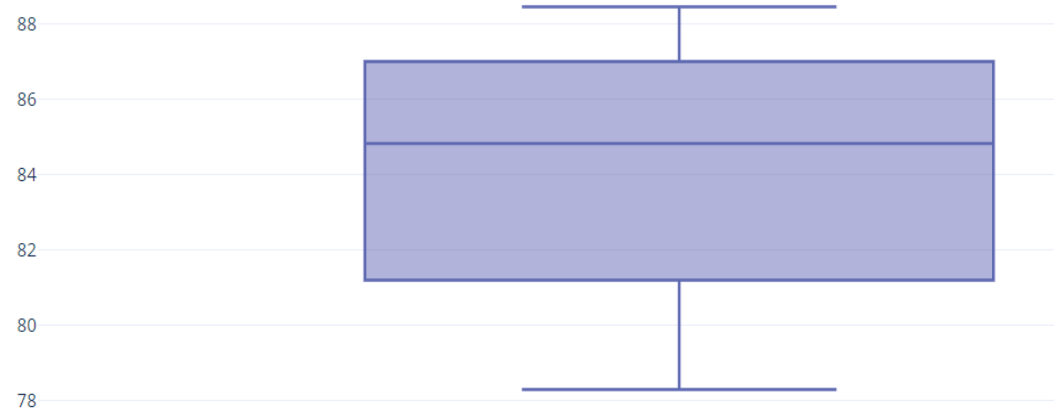

Figure 4.7: Accuracy Box plot for Support Vector Regressor

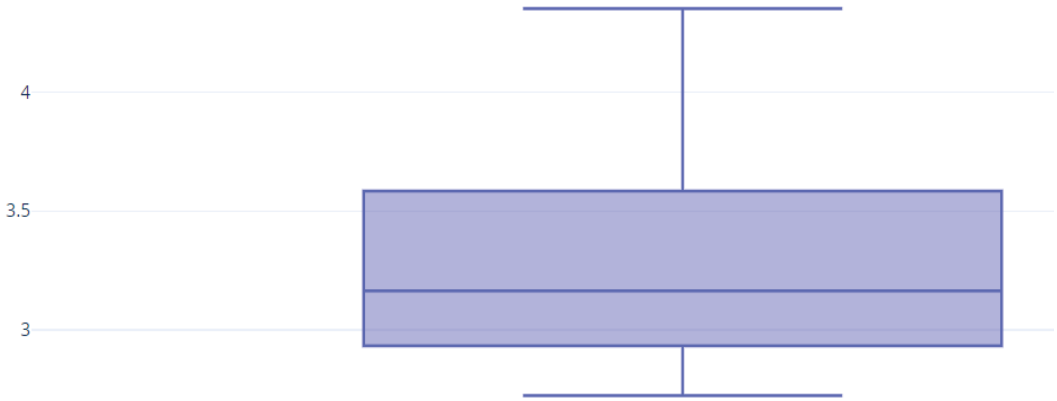

Figure 4.8: MAE Box plot for Support Vector Regressor

The box plot in Figure 4.6 shows the Max Error(ME) obtained by Gradient Boosting Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum ME of 0.464, the middle quartile represents a median ME of 0.441 and the lower quartile of the box plot represents a minimum ME of 0.425.

4.3 Support Vector Regressor

The box plot in Figure 4.7 shows the accuracy score(ACC) obtained by Support Vector Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum accuracy score of 88.45 percent, the middle quartile represents a median accuracy score of 84.82 percent and the lower quartile of the box plot represents a minimum accuracy score of 78.3 percent.

The box plot in Figure 4.8 shows the Mean Absolute Error(MAE) obtained by

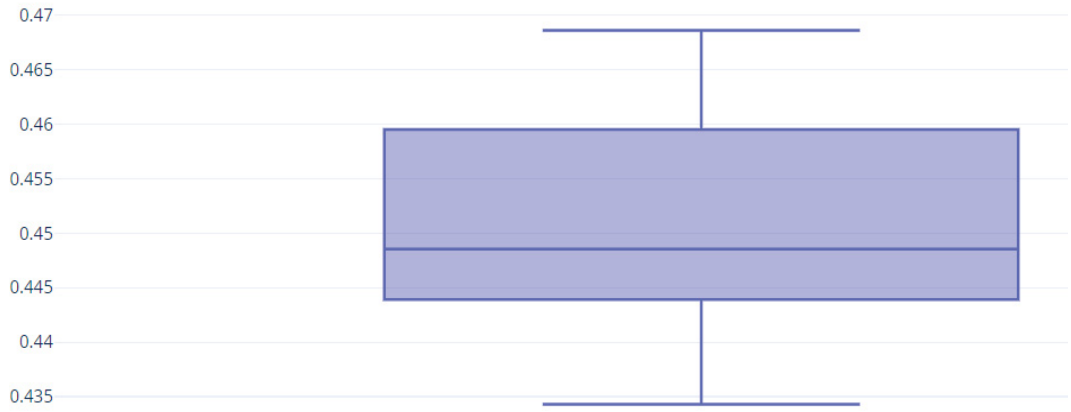

Figure 4.9: ME Box plot for Support Vector Regressor

The outcomes that Random Forest Regressor obtains are as follows:

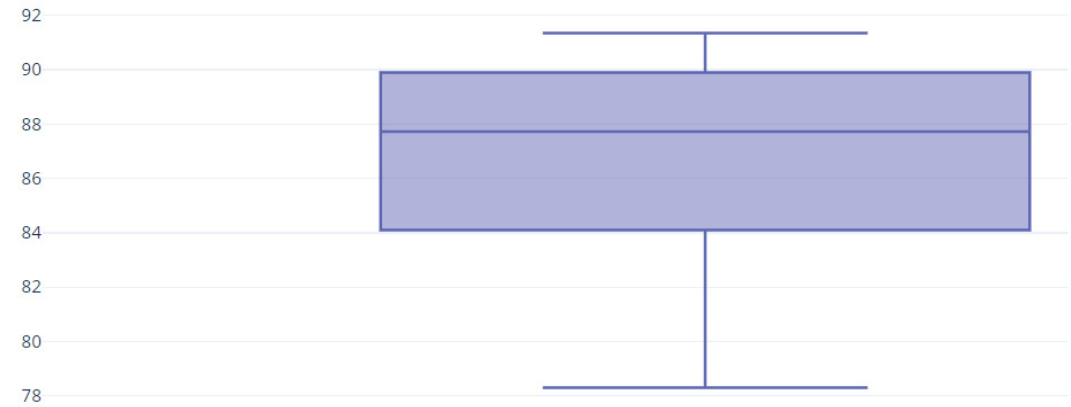

Figure 4.10: Accuracy Box plot for Random Forest Regressor

Support Vector Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum MAE of 4.3507, the middle quartile represents a median MAE of 3.1647 and the lower quartile of the box plot represents a minimum MAE of 2.7238.

The box plot in Figure 4.9 shows the Max Error(ME) obtained by Support Vector Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum ME of 0.4686, the middle quartile represents a median ME of 0.4485 and the lower quartile of the box plot represents a minimum ME of 0.4343.

4.4 Random Forest Regressor

The box plot in Figure 4.10 shows the accuracy score(ACC) obtained by Random Forest Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum accuracy score of 91.35 percent, the middle quartile

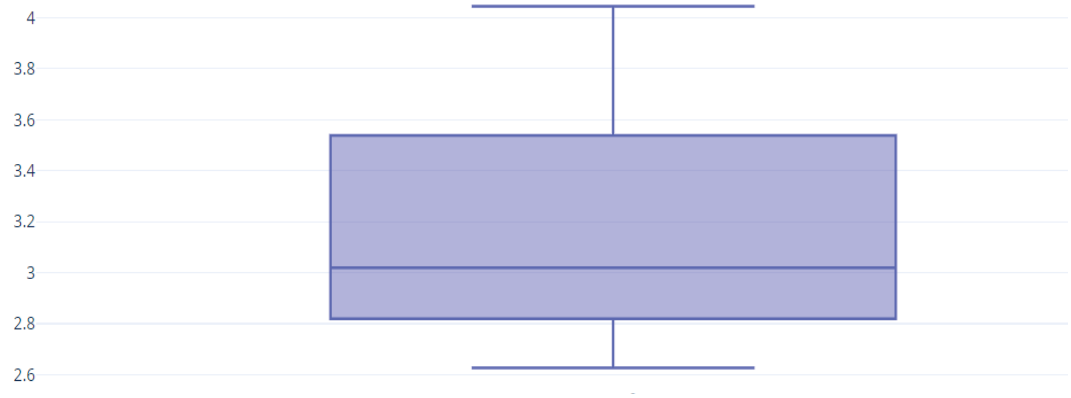

Figure 4.11: MAE Box plot for Random Forest Regressor

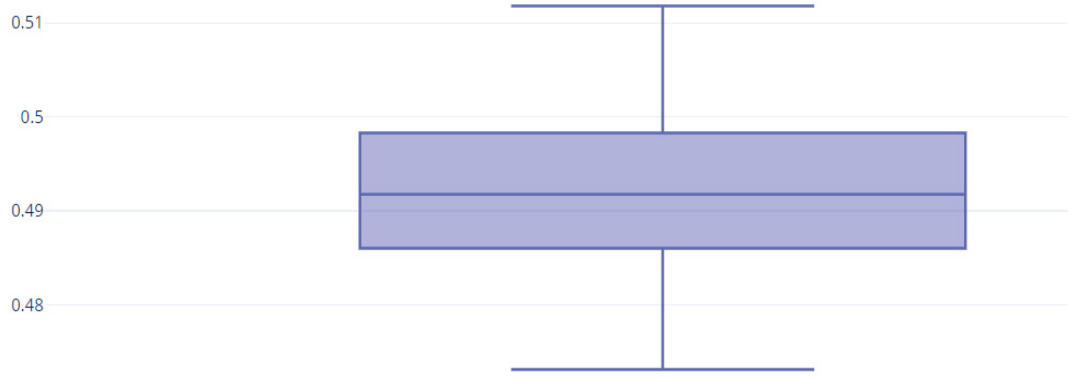

Figure 4.12: ME Box plot for Random Forest Regressor

represents a median accuracy score of 87.72 percent and the lower quartile of the box plot represents a minimum accuracy score of 78.31 percent.

The box plot in Figure 4.11 shows the Mean Absolute Error(MAE) obtained by Random Forest Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum MAE of 5.8058, the middle quartile represents a median MAE of 4.458 and the lower quartile of the box plot represents a minimum MAE of 3.4156.

The box plot in Figure 4.12 shows the Max Error(ME) obtained by Random Forest Regressor during a 10-fold stratified cross-validation test. The upper box plot represents the maximum ME of 0.6568, the middle quartile represents a median ME of 0.6135 and the lower quartile of the box plot represents a minimum ME of 0.5964.

4.5 Feature Importance

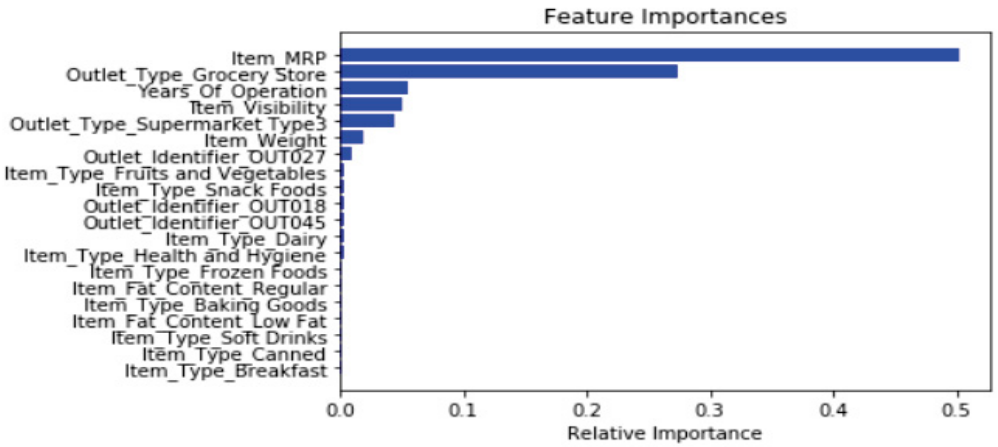

Figure 4.13: Feature Importance

Figure 4.13 shows that the feature importance of price of the products would depend primarily on the sales followed by the type of outlet and grocery store and the rest of the features would not even close to these features. There will surely be a huge impact on sales forecasting with these features.

4.6 Evaluation Results

Algorithms	Accuracy Score	Mean Absolute Error	Max Error
Random Forest Regression	87.72 percent	3.15	0.441
Gradient Boosting Regression	86.27 percent	3.198	0.491
Support Vector Machine	84.82 percent	3.21	0.448
Simple Linear Regression	81.2 percent	3.212	0.491

Table 4.1: Comparison of Evaluation Results

Table 4.1 shows the comparison of evaluation results where Random Forest Regression performed well with all the metrics accuracy score, Mean Absolute Error and Max Error. Random Forest Regression had the minimum error in predicting the sales when compared to the Simple Linear Regression, Gradient Boosting regression and Support Vector Machine. Simple Linear Regression demonstrated the worst performance with the highest error in all the metrics. A simplified tabular form based on the results is created above.

Chapter 5

Analysis and Discussion

5.1 Comparative analysis of Performance Metrics

In this section the average determined on the basis of ten iterations that were considered for all the metrics.

5.1.1 Average Accuracy Score

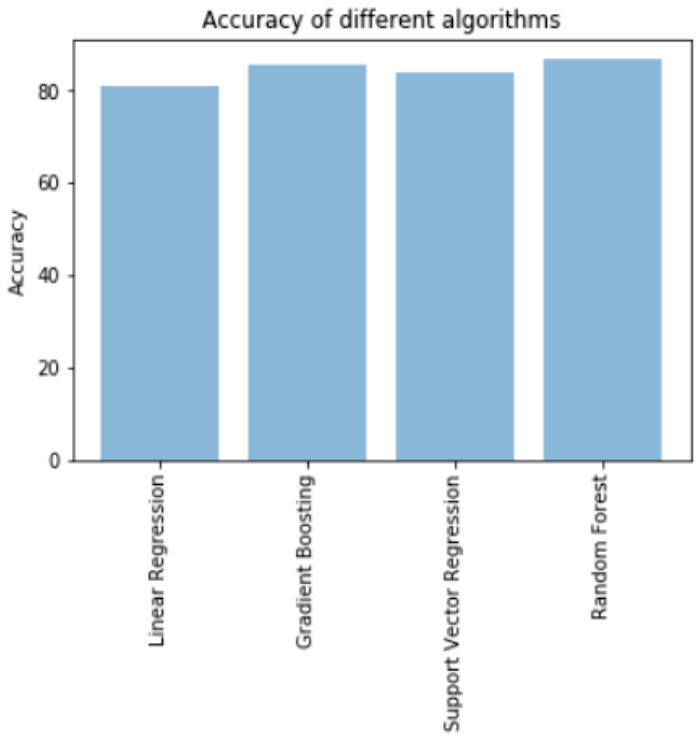

Figure 5.1: Average Accuracy score plot

Figure 5.1 shows the average accuracy score of the 10-fold stratified cross-validation obtained by the Simple Linear Regressor is 81.2 percent, followed by the Gradient Boosting Regressor with 86.27 percent accuracy score, then SVR with the 84.82 percent accuracy score and finally Random Forest Regressor with 87.72 percent accuracy

score. From figure 5.1, it can be seen that Random Forest Regressor is the best performer with approximately 88 percent accuracy score compared to other methods, and Simple Linear Regressor is the poor performer with an accuracy score of 81.2 percent.

5.1.2 Average Mean Absolute Error

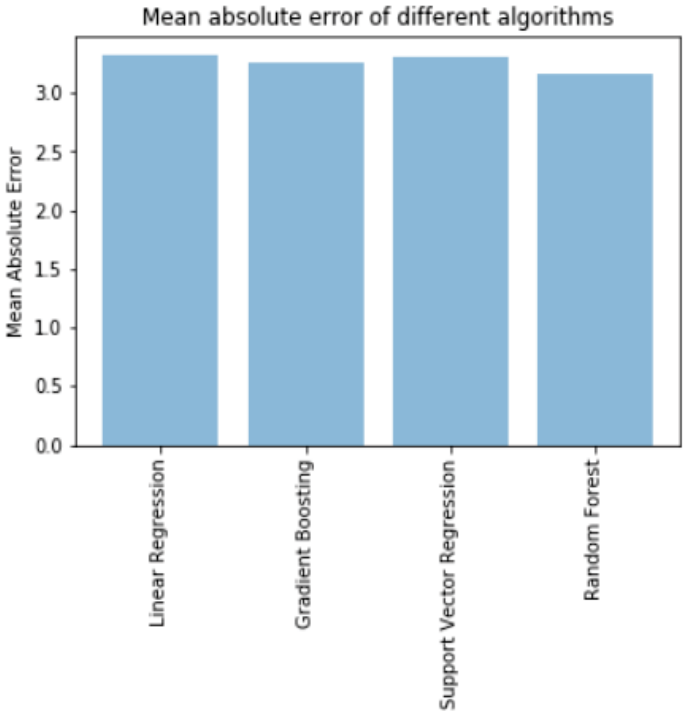

Figure 5.2: Average Mean Absolute Error plot

Figure 5.2 shows the average MAE of the 10-fold stratified cross-validation obtained by Simple Linear Regressor is 3.21 error, followed by Gradient Boosting Regressor with 3.19 error, then SVR with 3.21 error, and finally Random Forest Regressor with 3.15 error respectively. From figure 5.2, it can be shown that Random Forest Regressor is the best performer with less error relative to other approaches and with the highest error, Simple Linear Regressor is the poor performer.

5.1.3 Average Max Error

Figure 5.3 shows the average MAE of the 10-fold stratified cross-validation obtained by Simple Linear Regressor is 0.49 error, followed by Gradient Boosting Regressor with 0.491 error, then SVR with 0.448 error, and finally Random Forest Regressor with 0.441 error respectively. From figure 5.3, it can be shown that Random Forest Regressor is the best performer with less error relative to other approaches and with the highest error, Simple Linear Regressor is the poor performer.

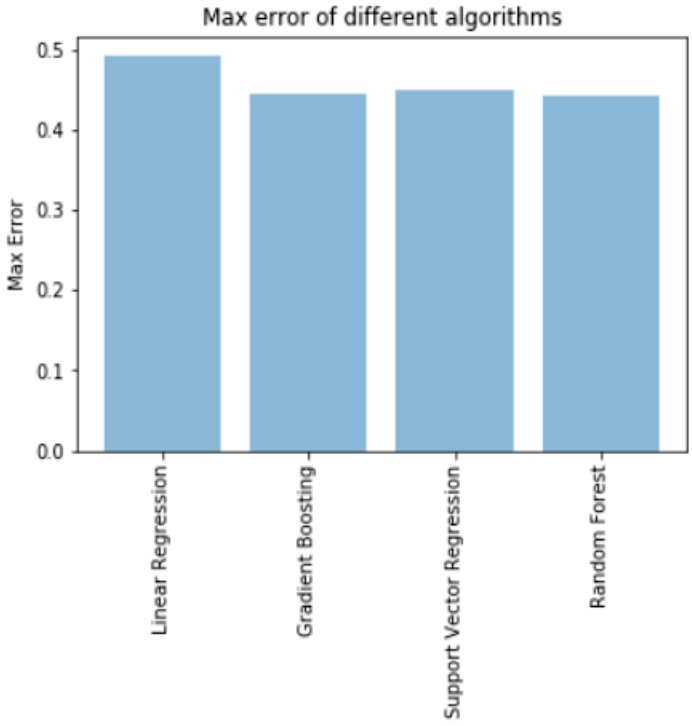

Figure 5.3: Average Max Error plot

5.2 Discussion

RQ1: What are the critical features that influence product sales?

Answer:

It was clearly observed in figure 4.13 that the price of the products and followed by the type of outlet and grocery store will heavily influence the product sales.

RQ2:

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

Answer:

Random Forest Regression is the most appropriate algorithm for forecasting the product sales. When compared to the Simple Linear Regression, Gradient Boosting Regression and Support Vector Regression, Random Forest Regression technique will produce the least error while predicting the product sales.

Average accuracy score, mean absolute error and max error for the Random Forest Regressor across the 10-fold stratified cross-validation is 87.72 percent, 3.15 and 0.44 error respectively which is quite impressive compared to other techniques. Simple linear Regressor produced the very poor results compared to the other techniques,

the average accuracy score, mean absolute error and max error across the 10-fold stratified cross-validation is 81.2 percent, 3.21, 0.49 error which is the poor one. And we can also observe from figure 4.13, Item price and outlet type grocery store are the critical features that will mainly influence the product sales. If the sales forecast is carried on every day across a large number of stores speed will play a key aspect in this process. Another useful metric to train the model, which will also play a crucial role while training several algorithms. The other important measure is the time required to train the model, which will also play a critical role while training different types of algorithms.

5.3 Contributions

As there are so many ongoing experiments that use statistical approaches and some traditional methods to focus on predicting item sales. Most researches have experimented by taking a single algorithm to predict sales. In this thesis Machine Learning algorithms such as Simple Linear Regression, Support Vector Regression, Gradient Boosting algorithm, and Random Forest Regression are considered for prediction and the most effective metrics such as accuracy, mean absolute error, and max error are considered for measuring algorithm efficiency. This method will be very beneficial in the future for advanced item sales forecasting.

5.4 Validity Threats

5.4.1 Internal Validity

Proper preprocessing of data will be done. There may be a high chance of less accuracy if one can not perform data preprocessing properly.

5.4.2 External Validity

The data and findings used for the experiments are justified and relevant. This problem can be solved if the performance metrics and algorithms are not chosen properly. Proper selection of algorithms and performance metrics will be done.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

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

6.2 Future Work

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

References

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

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

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

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

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

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

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

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

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

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

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

[12] Maike Krause-Traudes, Simon Scheider, Stefan Rüping, and Harald Meßner. Spatial data mining for retail sales forecasting. In *11th AGILE International Conference on Geographic Information Science*, pages 1–11, 2008.

[13] Stephen Marsland. *Machine learning: an algorithmic perspective*. CRC press, 2015.

[14] ML documentation. `https://www.mathworks.com/discovery/machine-learning.html)`. Accessed: 2020-04-22.

[15] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[16] Arvin Wen Tsui, Yu-Hsiang Chuang, and Hao-Hua Chu. Unsupervised learning for solving rss hardware variance problem in wifi localization. *Mobile Networks and Applications*, 14(5):677–691, 2009.

[17] Bohdan M Pavlyshenko. Machine-learning models for sales time series forecasting. *Data*, 4(1):15, 2019.

[18] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. *New advances in machine learning*, pages 19–48, 2010.

[19] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[20] Gradient Boosting documentation. `https://turi.com/learn/userguide/supervised-learning/boosted_trees_regression.html)`. Accessed: 2020-05-19.

[21] JN Hu, JJ Hu, HB Lin, XP Li, CL Jiang, XH Qiu, and WS Li. State-of-charge estimation for battery management system using optimized support vector machine for regression. *Journal of Power Sources*, 269:682–693, 2014.

[22] Wangchao Lou, Xiaoqing Wang, Fan Chen, Yixiao Chen, Bo Jiang, and Hua Zhang. Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian naive bayes. *PloS one*, 9(1), 2014.

[23] İrem İşlek and Şule Gündüz Öğüdücü. A retail demand forecasting model based on data mining techniques. In *2015 IEEE 24th International Symposium on Industrial Electronics (ISIE)*, pages 55–60. IEEE, 2015.

[24] Takashi Tanizaki, Tomohiro Hoshino, Takeshi Shimmura, and Takeshi Takenaka. Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, 79:679–683, 2019.

[25] Xu Ma, Yanshan Tian, Chu Luo, and Yuehui Zhang. Predicting future visitors of restaurants using big data. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 269–274. IEEE, 2018.

[26] Mikael Holmberg and Pontus Halldén. Machine learning for restaurant sales forecast, 2018.

[27] I-Fei Chen and Chi-Jie Lu. Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Computing and Applications*, 28(9):2633–2647, 2017.

[28] Malek Sarhani and Abdellatif El Afia. Intelligent system based support vector regression for supply chain demand forecasting. In *2014 Second World Conference on Complex Systems (WCCS)*, pages 79–83. IEEE, 2014.

[29] Jason Brownlee. *Introduction to time series forecasting with python: how to prepare data and develop models to predict the future*. Machine Learning Mastery, 2017.

[30] Python history. `https://en.wikipedia.org/wiki/Python_(programming_language)`. Accessed: 2020-04-29.

[31] Guido Van Rossum et al. Python programming language. In *USENIX annual technical conference*, volume 41, page 36, 2007.

[32] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[33] Wes McKinney. Pandas, python data analysis library. *see http://pandas. pydata. org*, 2015.

[34] Niyazi Ari and Makhamadsulton Ustazhanov. Matplotlib in python. In *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)*, pages 1–6. IEEE, 2014.

[35] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.

[36] Seaborn documentation. `https://seaborn.pydata.org/introduction.html)`. Accessed: 2020-04-26.

[37] Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, Cheng-Hong Yang, et al. Feature selection using pso-svm. *International Journal of Computer Science*, 2007.

[38] Tao Zhang, Tianqing Zhu, Ping Xiong, Huan Huo, Zahir Tari, and Wanlei Zhou. Correlated differential privacy: Feature selection in machine learning. *IEEE Transactions on Industrial Informatics*, 2019.

[39] Pearson documentation. `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)`. Accessed: 2020-04-25.

[40] Feature Importance documentation. `https://machinelearningmastery.com/calculate-feature-importance-with-python/#:~:text=Feature%20importance%20refers%20to%20a,feature%20when%20making%20a%20prediction.)`. Accessed: 2020-06-06.

[41] Kedar Potdar, Taher S Pardawala, and Chinmay D Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4):7–9, 2017.

[42] Cross validation documentation. https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85). Accessed: 2020-04-28.

[43] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[44] Accuracy documentation. https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085 Accessed: 2020-05-01.

[45] scilearn max error. https://scikit-learn.org/stable/modules/model_evaluation.html#max-error. Accessed: 2020-05-10.

[46] scilearn mean absolute error. https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error. Accessed: 2020-05-10.

Appendix A

Appendix

The graph below is a comparative analysis between the item sales and the item weight.

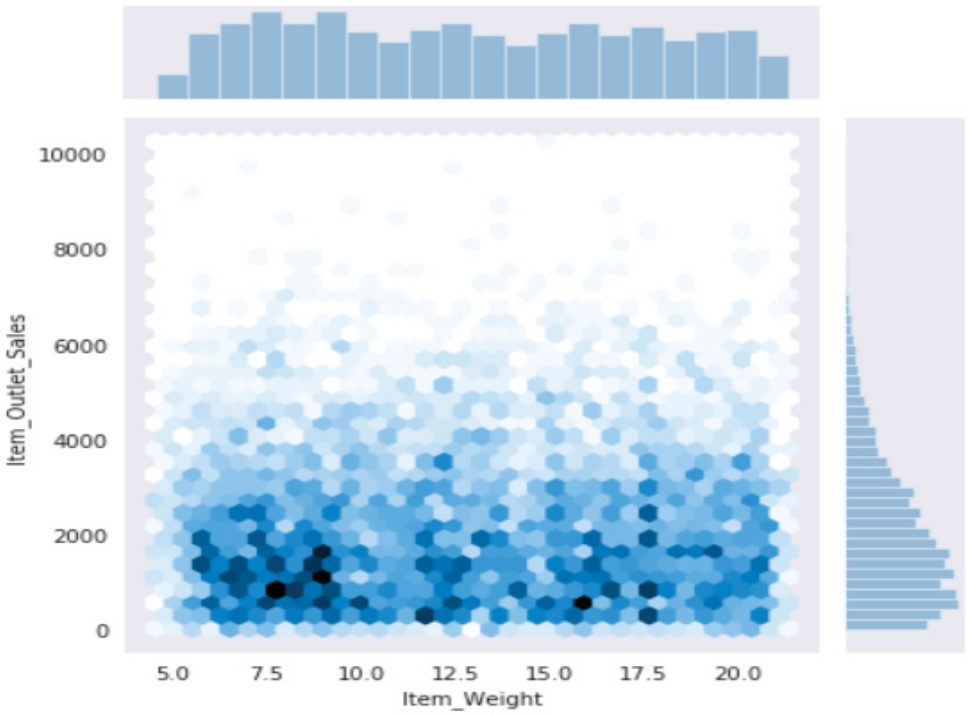

Figure A.1: Item outlet sales Vs Item weight

The graph below is a comparative analysis between the item sales and the item visibility.

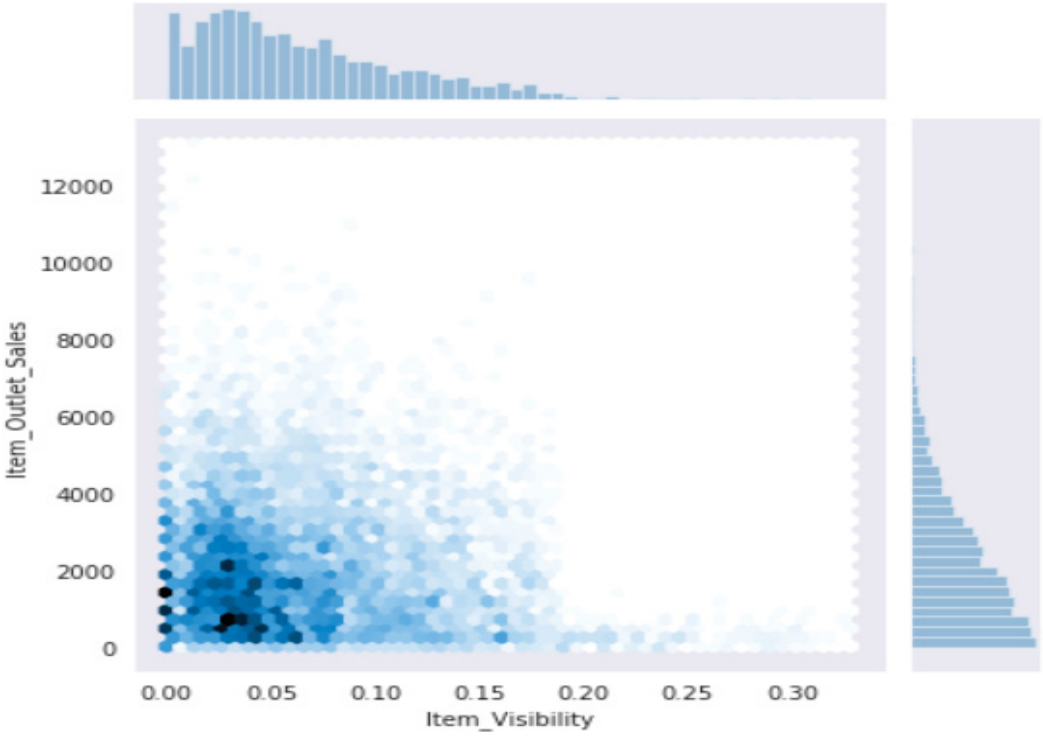

Figure A.2: Item outlet sales Vs Item Visibility

The graph below is a comparative analysis between the item sales and the item price.

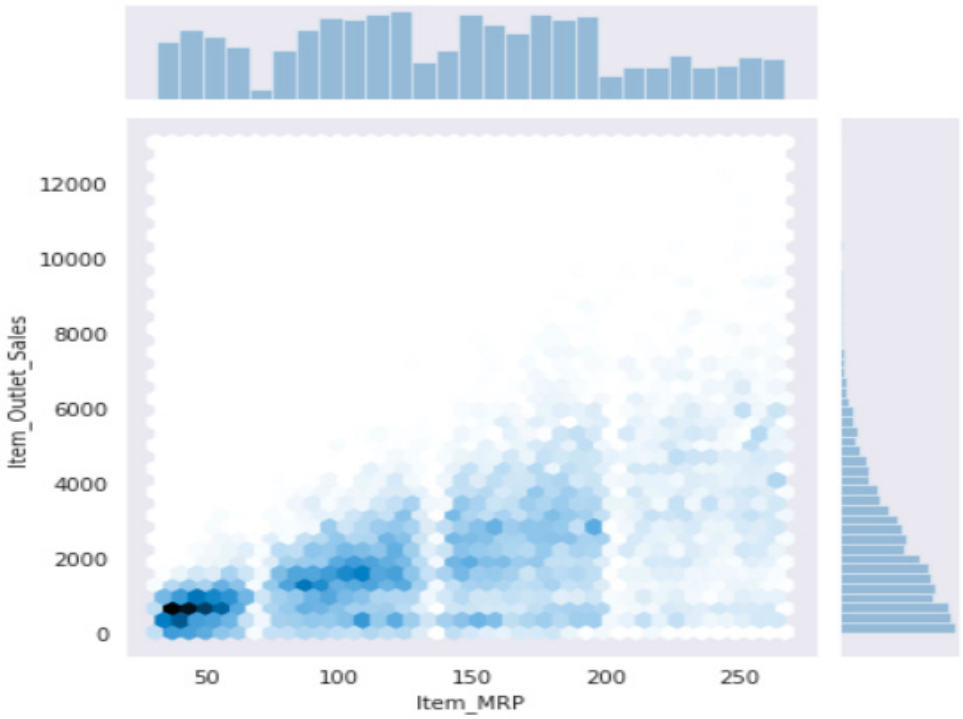

Figure A.3: Item outlet sales Vs Item MRP

The following graph illustrates the impact of item type on item outlet sales.

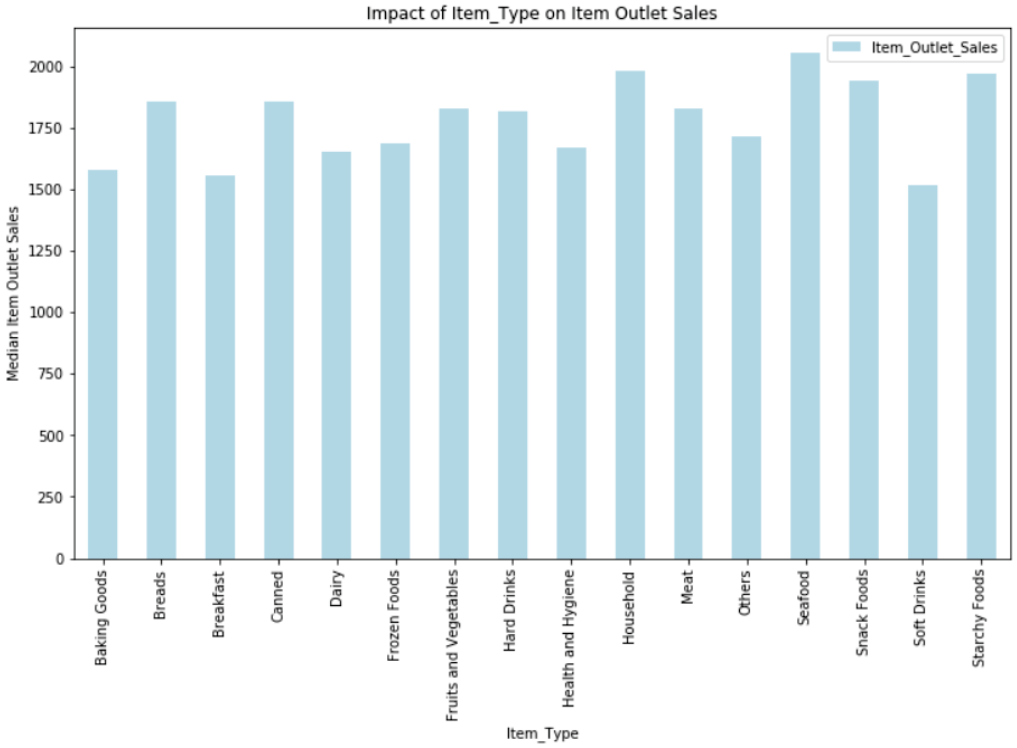

Figure A.4: Item Type Vs Median Item Outlet Sales

The following graph illustrates the impact of outlet establishment year on item outlet sales.

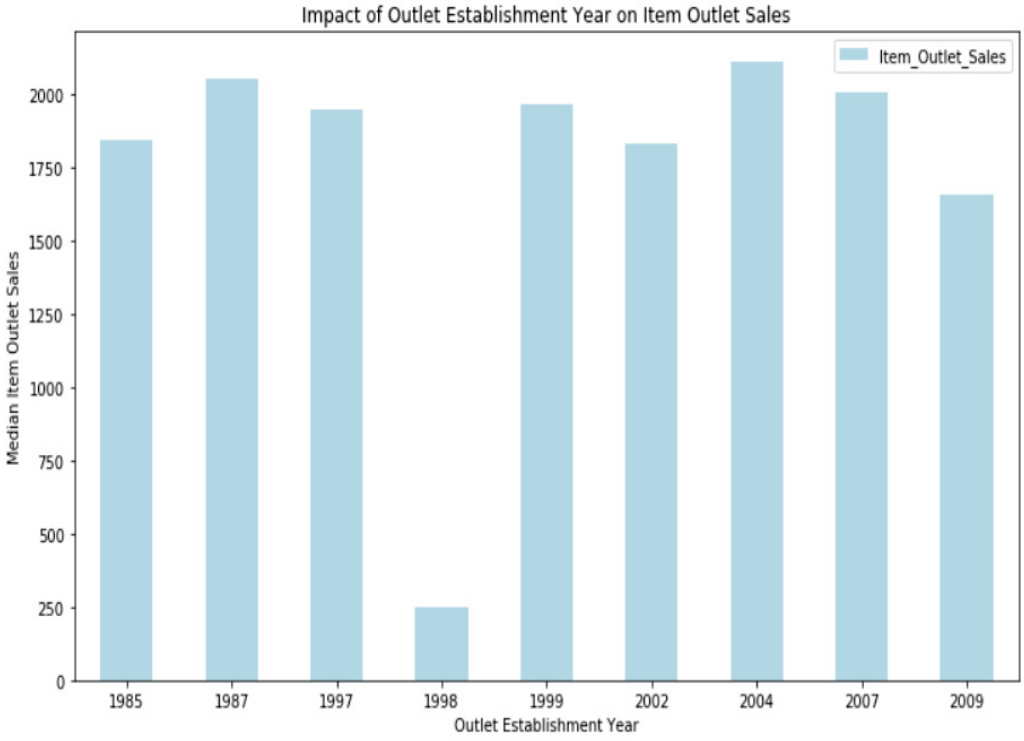

Figure A.5: Outlet Establishment Year Vs Median Item Outlet Sales

The following graph illustrates the impact of outlet sizes on item outlet sales.

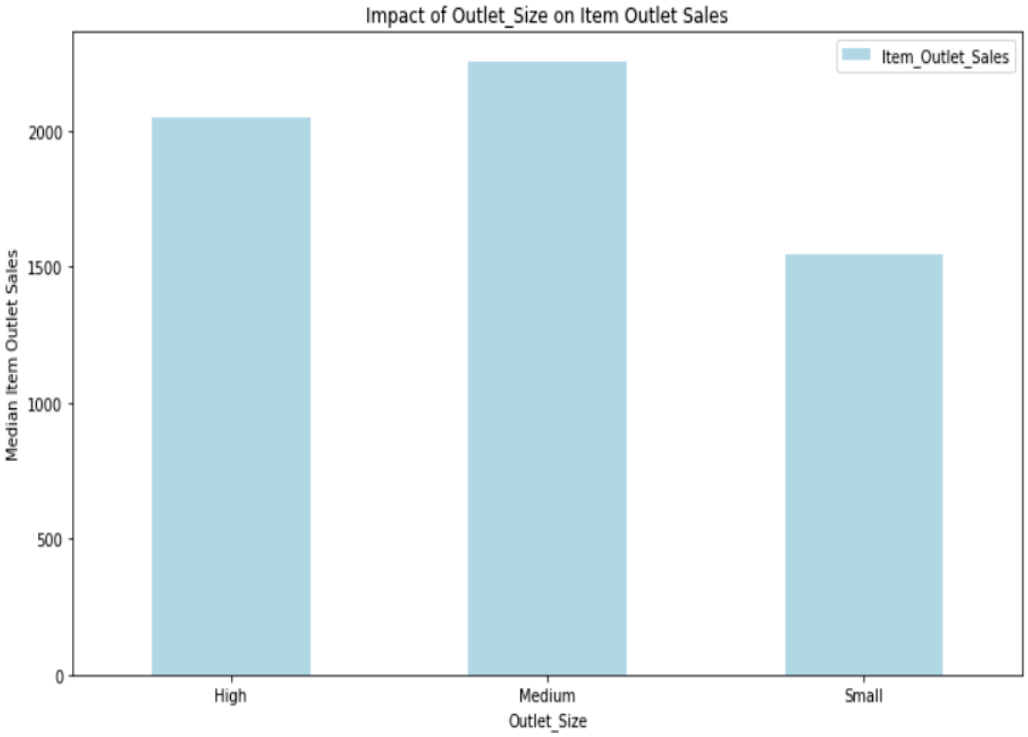

Figure A.6: Outlet Size Vs Median Item Outlet Sales

The following graph illustrates the impact of outlet location type on item outlet sales.

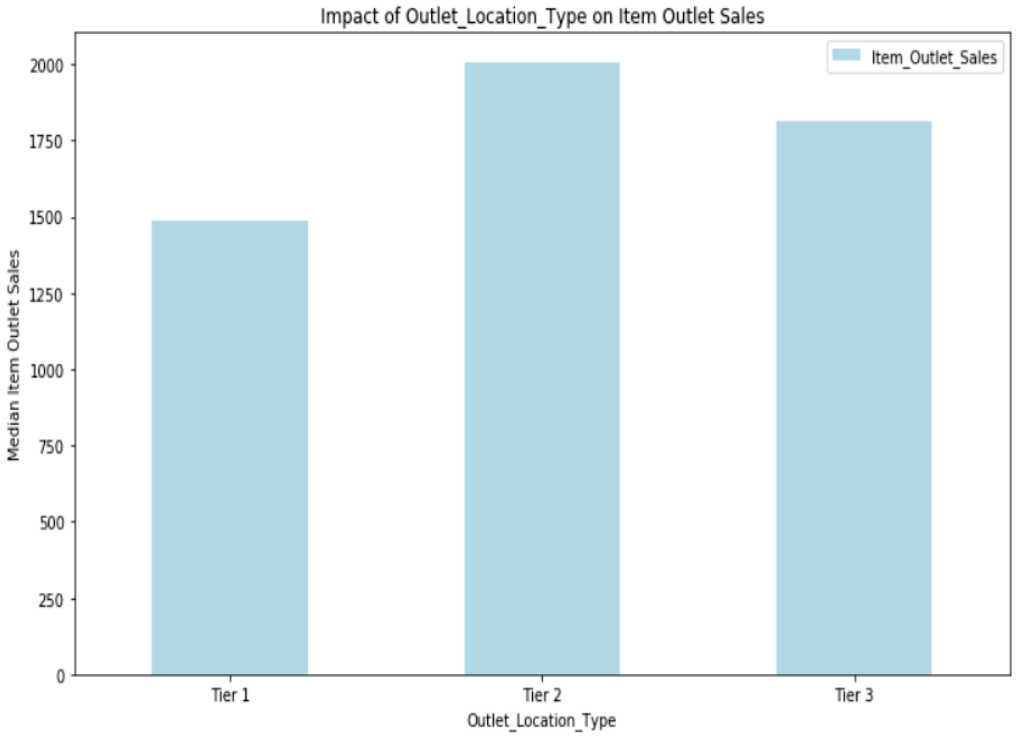

Figure A.7: Outlet Location Type Vs Median Item Outlet Sales

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden