Chapter 1

INTRODUCTION

Sales forecasting is the task of predicting or estimating sales revenue over a specific period of time. Forecasting sales is one of the most essential and decisive aspect of strategic planning. Proper sales forecasting facilitates companies to plan business requirements and prearrangements. In today's world, it is not possible to predict short time and long time performance and achievements without sales prediction. Sales forecasting also helps a company to administer its workforce, cash flow and resources. In the last decade, machine learning has grown to become a field with broad application potential and has gradually gained ground in sales forecasting.

For several reasons, while forecasting sales, retailers face many challenges. Season change, constant introduction of new items promotional activities plays the principal role. For eradicating these issues, retailers do sales prediction for accommodating large amount of sales data. It helps retailers to predict future customer behavior. Sales forecasting gives retailers the freedom of optimizing their revenue system by enabling more selective and suitable choices on promotion and pricing. Sales prediction also helps to predict the ups and downs of a country's economy. There are a wide range of applications where sales prediction can have an impact and prediction tools are used in many areas of business firms.

For our project, we will predict weekly sales based on historical data. The dataset contains weekly sales of various departments of similar types of different stores over different period of time. The project forecasts weekly sales based on different parameters of dataset following different methodologies. For our project, we accept this challenge and will try to correctly forecast sales.

1.1 Motivation of this Research

For making a place in this competitive world, it is very important for a company to predict sales. It is like a commitment for the sales department which must be achieves within a predicted time. Every company should decide in which department, they should emphasize more on expanding and which departments are the reasons for their loss. It is not possible to manage inventory without the solid idea of what future sales is going to be. For determining the best course of action, sales manager may analyze the trends at any time of fiscal year. Managers often need to analyze sales reports to identify market opportunities and areas where they could increase volume. For these reasons, I was motivated to do this project and decided sales forecasting as the research topic.

1.2 Objective of the System

The objective of the system is to predict weekly sales based on historical dataset of Walmart. Walmart is a reknowned departmental store in the United States of America. The dataset was provided by Kaggle. For this research, I have used data of six different stored where each store has 17 different departments.

The goal of the system are

i. To process dataset from different departments of different stores.
ii. To find out the weekly sales based on the historical data.
iii. To apply different algorithm to find out the accuracy and MAE.
iv. Finally, comparing the algorithms to find out which algorithm shows the most accurate result.

1.3 Purpose of Sales Forecasting

Every operation is carried out with the motive to fulfil some objectives. Sales Forecasting is also done for some objectives. The objectives of sales prediction are

- Predicting future customer behavior.
- Predicting the ups and downs of a country's economy.
- Saving business organizations from suffering financial loss.
- Facilitating companies to plan business requirements and prearrangements.
- Managing the flow of work of any organization.
- Facilitating organizations to decide where they should invest more and where investing can cause sufferings.

1.4 Research methodology

The methodology of this research is based on Linear Regression, Decision Tree Regression, Decision Tree Regression(max-depth=3) and Random Forest Regression. The dataset is divided into train data and test data. The accuracy of train data and test data is found out applying the algorithms and compared for finding out the most suitable algorithm. The used for forecasting sales of my project has relied on traditional statistical model. For predicting the outcomes of a future events, a machine learning model is exposed to data from which it learns patterns that are used to predict the outcomes.

1.5 Organization of the Report

We have formatted out report within five chapters from chapter 1 to chapter 5. Each of this chapter provides a detail description about the integration of the system. Each section deeply illustrates the concepts used in developing the system.

- Chapter 1 gives an overview of our project that we are embedded to develop. This chapter also provides a brief description about motivation of the research, goal of the system, objective of sales forecasting and methodology of this research.
- Chapter 2 provides background of this report. This chapter includes problem definition, a brief description of machine learning and machine learning scopes, methods of sales forecasting, types of sales forecasting and previous works that have been done to forecast sales.
- Chapter 3 includes the system development methodology that we have to follow for building our system. This chapter also outlines the techniques and algorithms that will be used in evaluating our systems.
- Chapter 4 is the implementation and result part which illustrates the collection of data and features of data, result of testing and makes a discussion.
- Chapter 5 draws the overall study and the future research that done later.

Chapter 2

BACKGROUND

In recent years, sales prediction has drawn much attention in many research communities as it helps in safeguarding economic and financial condition of an individual, a company as well as a country. Apart from this, sales forecasting also helps to adapt with the constant introduction of new products. Retailers can decide where they can invest without risk and more profitably by sales forecasting. Researchers are working on it from a long time ago and the struggle for coming up with the finest result did not still end.

Computers systems use machine learning effectively for performing a specific task and rely on pattern and inference instead of using explicit instructions. Machine learning can be called a subset of Artificial Intelligence. [2] For performing tasks, without explicitly programming, machine learning algorithms build a mathematical model of sample data known as training data.

For years, people have been forecasting weather patterns, economic and political events, sports outcome, health conditions and more. There are a wide variety of ways in which forecasts can be developed as we try to predict so many different events. [1] Using simple perceptivity, expert assumption and assessment or using past results to compare with universal analytical and time series technique small in number. There have been great improvement in accuracy of forecasting with the continual introduction of advanced data science and machine learning techniques. Forecasting is used extensively in business to make predictions such as demand, capacity, budgets, revenue etc.

2.1 Problem Definition

Sales forecasting is an essential and fundamental part of business management. Sales forecasting, also known as sales prediction helps to find out the potential issues when there is still time to avoid or mitigate them. It is not possible to manage inventory without the solid idea of what future sales is going to be. [3] Planning for growth and smooth cash flow is also maintained when a proper sales prediction is carried out. Sales forecasting is an estimation of sales volume that a company can expect to attain within a planned period. It is the projection of customer demand for the goods and services over a period of time.

There are some factors that effects the process of sales forecasting. Economic condition of the company and the consumers plays the most important role in sales forecasting. Past behavior of market, national income, disposable personal income, consuming habits of the customers also determines the ups and downs. Type and quality of product affects the estimation to a great extent. As markets are full of similar products manufactured by different firms, it is important for any company to make sure that their product maintains a good quality in the competition. It is obligatory to draw the attention of the customers through advertising. Companies also need to adapt with the changing technologies. [4] So, if a company keeps this factors in mind and do a proper planning, it can be expected that the company may not suffer from any financial and economic loss in the long run. Proper sales planning helps any company to make a space in the competitive world.

The importance of sales forecasting in any business organization is very high. A business organization can work systematically if they can predict sales properly. Forecasting enables project managers to set target of workers and fix responsibilities on every salesman. A company may result in wastage of resources if it loses its focus. Forecasting also helps to determine the production capacity.

2.2 Machine Learning

From the day computers were built, people have been looking for way to teach the purposes, hoping that someday they can program computers which will be able to improve their experience and can be smart by passing some experiments. We can imagine the day when computers can replace human by their human like intelligence. Machine can find techniques and solutions to solve problems effectively based on relative data.

Machine Learning is a core subarea of Artificial Intelligence where systems can learn data without being explicitly programmed. For example, a machine learning system can learn e-mail receiving and differentiate spam and non-spam messages from each other. Certainly, it can be said that the topic of machine learning plays a highly significant role in the field of computer science and game technology. [12] Machine learning is used in many industries for applications in banking and financial sectors, healthcare, retail, publishing and social media etc.

2.2.1 Examples of Machine Learning Problems

There are so many types of machine learning problems. Many of these are on classification and regression problems. Some examples of such problems are :

- **Optical Character Recognition** : Categorize image of handwritten characters by the letters represented.
- **Face Detection** : Detect faces in images or indicate if a face is present.
- **Spam Filtering** : Identify email messages if it's a spam or not.
- **Medical Diagnosis** : Diagnose a patient as a sufferer or non-sufferer of any disease.
- **Customer Segmentation** : Predict which customer will respond to a particular promotion.
- **Fraud Detection** : For identifying fraudulent, identifies credit card transactions.
- **Weather Prediction** : Find the weather of near future like whether it will rain tomorrow or not.
- **Sales Prediction** : Predict the amount of sales of any organization.

Machine learning means that how computers can learn and improve their performance based on data. It is a research area for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data. For example, while playing games with machine, we can see that it can make decisions towards the winning and by several playing experiences, it can make better decisions than previous one.

2.2.2 Types of Machine Learning Problems

Different types of machine learning problems are discussed below –

2.2.2.1 Supervised Learning

In supervised learning, the aim is to learn a mapping from the input to an output whose correct values are provided by the supervisor. At first, an algorithm is trained with the training data and it creates a model which is used as the supervisor. Here, labeled data is used. So, if the test data is given as input, then using the model, the unknown class level of data can be found. Supervised learning problems can further grouped into classification and regression problems.

- **Classification** – For classification, only two class labels are used : yes and no. For example, if a dataset has several attributes defining patient's physical condition and the class level indicates if the patient has diabetic or not.

- **Regression** – Where class level can be of any real values and there are more than two class labels. For example, if the class label of a dataset started above indicates the disease name, then it will be addressed as regression problem if there are more than two class labels like a patient has diabetic, a patient does not have diabetic and a patient has risk that he can be a diabetic patient in near future.

- **Anomaly Detection -** Anomaly detection is identifying simply unusual data points. In fraud detection, the approach that anomaly detection takes is to simply learn what normal activity looks like and identify anything that is significantly different. For example, any unauthorized access of bank account or credit card service without permission can be detected.

2.2.2.2 Unsupervised Learning

In unsupervised learning, there is no supervisor only the input data is given. The aim is to find regularities and similarities in input. Here, input data is not labeled also. Unsupervised learning problems can be grouped into clustering and association problem.

- **Clustering** – A clustering problem is such a problem where we want to discover the inherent grouping in the data. From these types of algorithms, several clusters can be produced where maximum similarity in some cluster elements and maximum dissimilarity among different cluster elements are found. For example, if we consider clustering in a dataset we can keep the data with similar properties in one cluster and data with different properties in different clusters. Thus maximum similarity among the element in same cluster and maximum dissimilarity among different clusters can be maintained.

- **Association** - An association problem is such a problem where we want to discover the associativity among the elements. For example, a customer who buys short has maximum possibility to buy pants. So, shirt and pants have association between them.

2.2.2.3 Semi-supervised Learning

Semi-supervised learning is a class of machine learning technique that make use of both labeled and unlabeled examples while learning a model. In one approach, labeled models are used to learn class models and unlabeled examples are used to refine the boundaries between classes.

2.3 Methods of Sales forecasting

Sales forecasting can be carried out by different methods. Among them, four methods are discussed below :

2.3.1 Sales Force Composite Method

When sales representatives of the business are asked to carry out sales forecasting of their respective area, it is called sales force composite method. It is a bottom up approach where sales force gives their opinion on sales trend to the top management. The advantage of this method is that it is a practical approach and each sales territory gets focus. Because of large production sample, the method is more reliable. So, the success of this method depends on the efficiency of the sales people. [5] For carrying out proper forecasting, each salesman must be committal and should be careful that they always give correct information. As the company's sales forecast is made on this basis of the forecast for territorial sales, the method can be considered as a grass-root level method.

2.3.2 Jury of Executive Opinion Method

By this method, sales forecasts is made based on the top executives of the company. The jury method can be reliable only if all the executives are hardworking and well informed about the economic condition of the company. For arriving at a conclusion, detailed analysis is not required. The predictions made by this method is not based on facts or figures. [6] The executives describe the past performance of the company, observes the present market condition carefully and then, predicts for future sales of the company. So, the jury members become responsible for any achievement or loss in the company.

2.3.3 Customers' Expectation Method

The customers are directly approached and based on their requirements, the future is ascertained. This method is best suitable for those industries where the number of customers are limited and the customers can be personally contacted. This method works best for industrial goods marketing. Customers' expectation method may not be suitable for consumer goods marketing. Long term plans may not be that much successful but short term plans have an impactful outcome. The method should be conducted in such a way that company can adapt when there is any change in customers' expectation. So, each company should also try to predict what changes can come in customers' expectation and what steps need to be taken for fulfilling customers' never ending expectation.

2.3.4 Statistical Method

Statistical method uses statistics based on historical data to predict what could happen out in the future. Statistical Methods for Forecasting is a complete, readable treatment of statistical models. Statistical method is generally used to produce short-term forecasts. For creating statistical forecast, complex mathematical terms are used. For that reason, most companies rely on advanced software to accomplish this task.

For any statistical method of forecasting, the very first important task is to collect raw data. After that, the organization of data is very important for coming to an appropriate conclusion. This method depends on the type of experiment and the desired output of the experiment. [7]

For statistical treatment of data, classifying data into commonly known pattern is very important. Data should also be described properly for statistical forecasting. For our sales forecasting, we will follow statistical method of forecasting. We will collect raw data of previous sales of a departmental store and predict the future weekly sales based on the collected data.

2.4 Types of sales forecasting

Based on the period needed, forecasting can be categorized into three types. This types are briefly described below :

2.4.1 Short-run Forecasting

Short-run forecasting covers maximum a year or it can be half-yearly, quarterly, monthly or even weekly. This type of forecasting is usually made for tactical reason and forecasting includes production planning and control. Short time cash requirements that need to be made for seasonal sales are also included in this type of forecasting. [8]

Short-run forecasting is suitable for fast moving factors like providing working capital, establishing sales quotas etc. This type of forecasting reduces the cost of raw materials and machinery. If the period of forecasting is short, it is easy to have proper control of inventory and setting sales target is also easy.

2.4.2 Medium-run Forecasting

Medium-run forecasting covers from more than one year to two or four years. This type of forecasting is made for minor strategic decisions in connection with the operation of the business. In medium-run forecasting, it is easy for management to control budgets, expenditure, production etc. If the forecasting turns out to be over-optimistic, it can cause damage to the budget and resources and there is possibility that the organization will be left with unsold stock. For machineries to be purchased to meet increased production, for short term capital requirements, for achieving expected sales, medium-run forecasting is carried out.

2.4.3 Long-run Forecasting

Long-run forecasting may cover up to twenty years or more. Forecasting might be needed for a decade or more for some heavily strategic, capitalized industries such as ship-building, petroleum refinery, paper making industries etc. [10] This type of forecasting is carried out for major strategic decisions that are needed to be taken within an organization. For any business organization, accuracy and significant time to onset is more vital than speedy updates. [11] Long-run forecasting relies more in government policy, social change and technological change. Economic depression, population changes and change in competition are also taken into account for long-run forecasting. [9]

Sales forecasting trends in long term is a different story as accurate forecasts allow a business to position top in the industry. As panning is done a long time ago, advance notice gives the business to change in plan and implement new strategy. Long term forecasting needs a deep knowledge of the inner workings of the industry.

For our project of forecasting sales, we are following long time forecasting. As we are using the historical data of long time and by using the data, we are planning for a long period of future, we need to follow the requirements and planning process of long time forecasts. Sales forecasting deals with long time requirements of products and demands of customers. We had planned for the new unit of production or expansion of existing unit to meet the demand. So, for sales forecasting, it is obligatory train the personnel so that man-power requirement can be met in future.

2.5 Related works of sales forecasting

Forecasting sales is a common task performed by any organization. For years, different organizations are forecasting sales for making a place in the competitive world. With the introduction of new technologies, it is becoming more and more difficult to make good use of new machineries and meet the never ending requirements of the customers. So, organization must plan for their sales and implement the plans for becoming successful. Some project on machine learning for sales prediction is discussed in this part of the report.

SVM is used as a prediction tool for forecasting sales in food industry and the system was proposed by Levis et al. [13] He predicted monthly sales of a single product 12 months ahead. In his research, the Mean Average Percentage Error was 7%. The authors suggested that neural network can be used as an alternative of SVM. In his food sales forecasting, the conclusion was also supported by Pillo at al [14] who found out that RBFN outperforms SVM for daily sales prediction. He trained one year data and SVM scored the lowest average MAPE (61%). Krause Traudes et al also successfully used SVM for predicting total sales of a store on aggregate level. [15]

A medical related sales forecasting of a drug store was done by Hongyu Xiong , Xi Wu and Jingying Yue. [16] They used AR model to predict the sales with small discrepancy to the test data and RF and SVR to find relations between store mean sales and other features.

A clustering based sales forecasting was done by Support Vector Regression. [17] The system used k-means algorithm to first divide training dataset into several disjoint clusters. Then SVR is applied in each group to forecast sales. The result of the experiment revealed that clustering based sales forecasts outperforms SVR and is an effective alternative of computer server sales forecasting.

For automobile market sales forecasting, a prediction was carried out [18] by using SVM, DT, KNN, RF, OLS and QR algorithms. OLS and QR delivered comparatively poor results. They predicted yearly, quarterly and monthly sales based on historical dataset.

A prediction on car sales was carried out by James barkovec. [19] The research developed a short-run general equilibrium model of the automobile market by combining a discrete choice model of consumer automobile demand with simple models of new automobile production and used vehicle scrappage.

A prediction on Sales Forecasting in Fashion and Apparel Industry was carried out by Oun Abbas, Khurram Shahzad, Ghulam Ali, Umar Sarwar. [20] Their model helped the suppliers to control the production volume of apparel products and reduced the un-wanted stock. For project, raw data was first collected and then preprocessed. Time series analysis has been done to build the model. The research used many techniques such as Regression, Seasonal Index and Quarter Moving Average technique.

Apart from that, many research we carried out to predict house sales. A research was carried out where the price is predicted. [21] The application was implemented as a regression problem that tried to estimate the market price of a house given features retrieved from public online listings. For the research, several machine learning techniques were applied including regression trees, k-nearest neighbors, support vector machines and neural networks, identifying advantages and handicaps of each of them. In the research, the regression showed the best result. The research found the smallest mean absolute error is 338,715 euros, and the best median absolute error is 94,850 euros.

For our project, we are using the historical dataset of Walmart which is a famous departmental store in the United States of America. The dataset was provided by Kaggle. We will apply different machine learning techniques such as Linear Regression, Random Forest etc. We will test the accuracy, MAE and MSE of the dataset and will try to find out which techniques fits best.

Chapter 3

METHODOLOGICAL APPROACH

To complete the thesis work, we need to approach according to a plan. As sales forecasting is a problem that can be solved applying machine learning techniques, we need to research different techniques and try to find the perfect method that gives the best result. At first, we need to know details about what is the problem actually that we need to solve and then we need to decide how to are going to solve the problem. So, in this part of the report, we discuss about the whole sales forecasting process and the ways to forecast sales.

For our project, at first we will split out dataset into training data and test data using test_train_split. The training data will be used as the supervisor of test data. We will fit training data into different data mining algorithm and then predict weekly for test data. Then we will find accuracy, mean absolute error and mean squared error of all the algorithms. At last, we will compare the results of the algorithms and will try to find out which algorithm fits the process best. A brief description of our dataset is also added in this chapter.

In this part of the report, we will discuss the methodological for building the sales forecasting system.

3.1 Methodology of Machine Learning Techniques

The figure below shows the steps needed to carry out any model building operation using machine learning techniques.

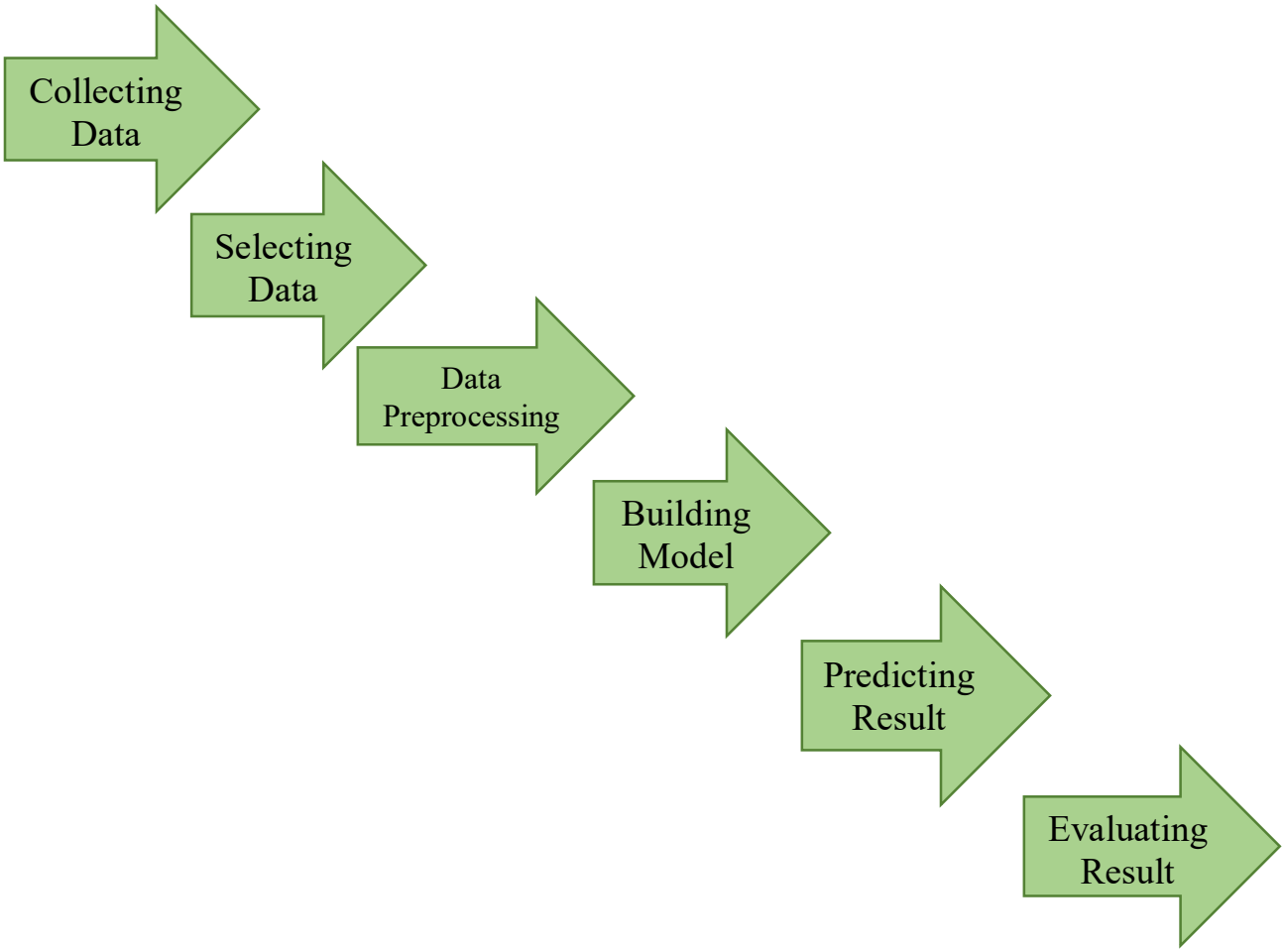

Fig 3.1 : Methodology of Machine Learning Technique

3.2 Methodological Approach of Sales Forecasting System

The figure below shows the methodological approach of sales forecasting system.

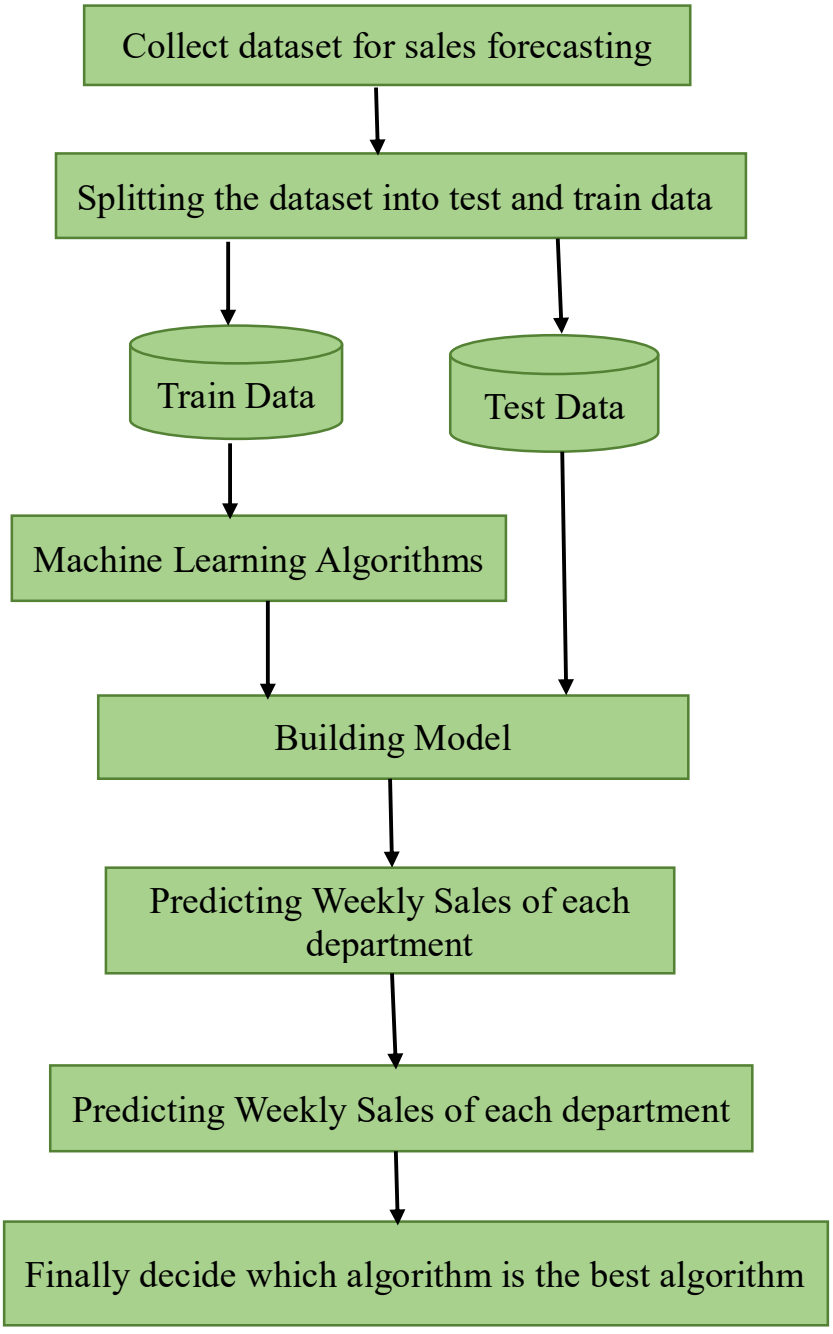

Fig 3.2: Methodological approach of sales forecasting system

For sales forecasting system, at first we have collected data from a famous repository named kaggle. We did some preprocessing on the data. Then we have split our into test data and train data. We have kept two third data as training data and one third data as test data. After that, we have fitted our dataset into different regression algorithm and built a model. Then comparing with the model, we have predicted weekly sales for each of the departments of the stores.

After predicting weekly sales, we followed some steps for deciding which algorithm fits our system the most. We have used different regression algorithm and tested Accuracy, Mean Squared Error (MSE) and Mean Absolute Error (MAE). The best algorithm will have the maximum accuracy and minimum mean squared error and mean absolute error. So, based on this characteristics, we have decided which one is the best algorithm.

3.3 Algorithms

There are several pre-developed algorithms that are used in machine learning to fed training data and produce an acceptable model. Machine learning provides huge number of algorithms to solve different types of problems. For out project, we are using the below mentioned algorithms :

- Decision Tree
- Linear Regression
- Random Forest

In this part of report, we will briefly discuss the algorithms.

3.3.1 Decision Tree

A decision tree is a decision support tool. A tree like graph is used or model of decisions and their possible consequences are used by decision tree. It is one way to display an algorithm. In a decision tree, each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test and each leaf node or terminal node holds a class label. A decision tree has a topmost node which is the root node.

Decision tree has become very popular machine learning algorithm because of some reason.

- Construction of a decision tree classifiers does not need any domain knowledge or parameter settings, and therefore, is appropriate for exploratory knowledge discovery.
- Decision tree can handle multidimensional data.
- The representation of acquiring knowledge in tree form is intuitive and generally easy to assimilate by humans.
- The learning and classification steps of decision tree induction is simple and fast.
- It allows the addition of new possible scenarios, which makes it convenient.

With vast popularity, decision tree has some disadvantages.

- Firstly, for data including categorical variables with different number of levels, sometimes decision trees are biased in favor of those attributes with more levels.
- Secondly, in constructing a decision tree, the dataset is repeatedly divided into subtrees, guided by the best combination of variables. However, finding the right combination of variables can be difficult.
- Thirdly, a decision tree constructed based on a small sample might not be generalized to future, large samples.

So, if a dataset down not have these characteristics, decision tree algorithm can be applied on the dataset to acquire good result.

3.3.2 Linear Regression

Linear regression is commonly used for predictive analysis. Linear regression relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Linear regression examines two things. They are

- does a set of predictor variables do a satisfying job in forecasting an outcome (dependent) variable?
- Which variables in appropriate are convincing predictors of the outcome variable, and in what way do they–pinpointed by the magnitude and sign of the beta estimates–impact the outcome variable?

Linear regression looks for statistical relationship, but not for deterministic relationship. Two variable are said to be in a deterministic relationship if one variable can be accurately expressed by the other. For example, using height in feet it is possible to accurately predict meter. For determining relationship of two variables, statistical relationship is not accurate.

The major uses of linear regression are :

- Determining the strength of predictors
- Forecasting an effect
- Trend forecasting

3.3.3 Random Forest

Random forest or random decision forests are an ensemble learning approach for regression, classification and other tasks. Random forest accomplishes tasks by building a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It overcomes many limitations of decision tree including generalizable problem and gives more accuracy than decision trees.

In this section, the working strategy of decision trees will be described. From a dataset, many decision trees can be built if at each node, attributes are selected randomly or from a random subset. As a result, many models can be found. From those models, prediction can be combine by majority voting or y taking averages. As a random forest in an ensemble of multiple decision trees, it is often more accurate than any individual decision tree. This is because each individual model has its own strength and weakness in predicting certain outputs. Suppose, an algorithm gives wrong output in a special case but others can give correct result on that case. So random forest will show the right result. Thus more accuracy is achieved. Besides, decision trees are constructed randomly for random forest. So, it overcomes bayesness problem to a great extend.

With all of these advantages, random forest has some limitations too. Random forests are considered “black-boxes”, because they comprise randomly generate decision trees and are not guided by explicitly guidelines in predictions.

Chapter 4

IMPLEMENTATION AND RESULT

The results of our project are weekly sales, accuracy, mean squared error and mean absolute error. The training dataset and test dataset will vary with each implementation as algorithms will randomly take test data and training data. So, the accuracy, mean squared error and mean absolute error will also change.

4.1 Analysis of data

Analysis of data includes collection of data, description of data, dictionary details of data and count plots of attributes of the dataset.

4.1.1 Collection of data

Sales forecasting by machine learning algorithms has been done many times by different researchers. Different dataset for sales forecasting is available in different resources. The dataset we are using for our project is provided by kaggle. [22] The dataset is of a famous departmental store of United States of America called Walmart. The dataset we are using has fifteen different attributes and 13728 rows.

4.1.2 Description of Dataset

Collection of data is called dataset. The dataset is described in the table below.

Table 4.1 : Description of data

Features	Description
Store	The number of store. There are 7 different stores.
Dept	The number of department. Each store has seventeen departments.
Date	The date of the particular day.
Store_size	The size of the store.
Temperature	Average temperature of the region.
IsHoliday	Whether the day is a holiday or not.
Fuel_price	Cost of fuel in the specific region.
Markdown 1-5	anonymized data related to promotional markdowns that Walmart is running. Markdown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
CPI	Consumer price index.
Unemployment	The rate of unemployment.
Weekly_Sales	Weekly sales for the given department in the given store.

4.1.3 Dictionary details of dataset

The dictionary details of the dataset used for forecasting sales are listed below :

Table 4.2 : Dictionary details of data

Features	Data Type	Missing Value	Unique Value	Count
Store	int64	0	6	13728
Dept	int64	0	16	13728
Date	Object	0	143	13728
Store_size	int64	0	6	13728
IsHoliday	Bool	0	2	13728
Temperature	float64	0	796	13728
Fuel_price	float64	0	248	13728
Markdown 1	float64	8802	306	4926
Markdown 2	float64	9755	235	3973
Markdown 3	float64	9139	271	4589
Markdown 4	float64	8882	300	4846
Markdown 5	float64	8802	306	4926
CPI	float64	0	858	13728
Unemployment	float64	0	72	13728
Weekly_Sales	float64	0	13701	13728

4.1.4 Count plot of attributes

A count plot is a histogram that shows the count of observations of each categorical bin using bars. The count plot of some attributes of the dataset is shown in this part.

One of the attribute of the dataset is store which has six different values. The count plot of store is shown below :

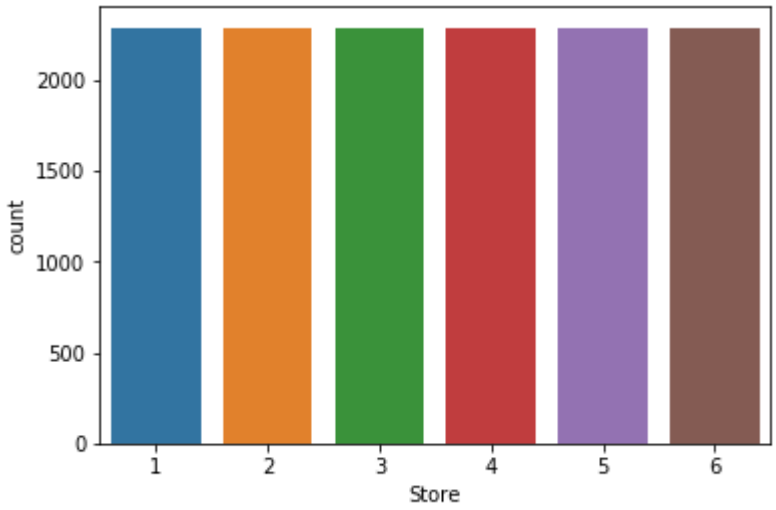

Fig 4.1 : Count plot of Store

Another attribute is department which has seventeen different values. The count plot of department is shown below :

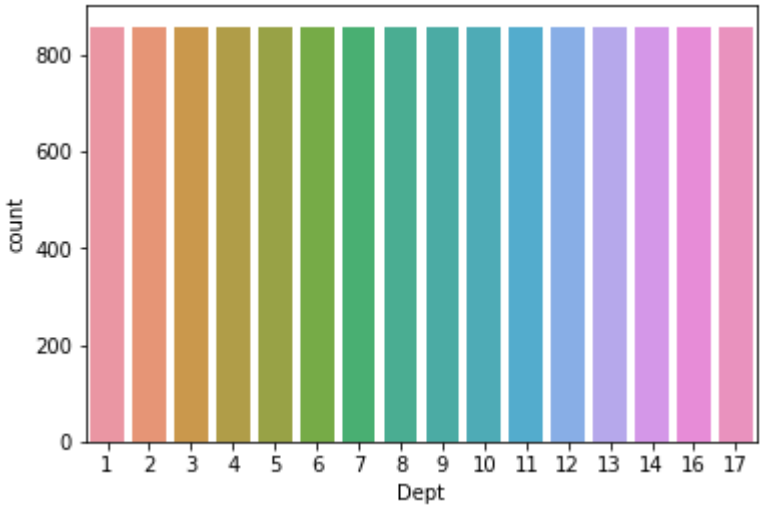

Fig 4.2 : Count plot of Department

Whether a specific day is a holiday or not is shown in the attribute IsHoliday. The specific attribute has Boolean values True and False. The count plot is shown below:

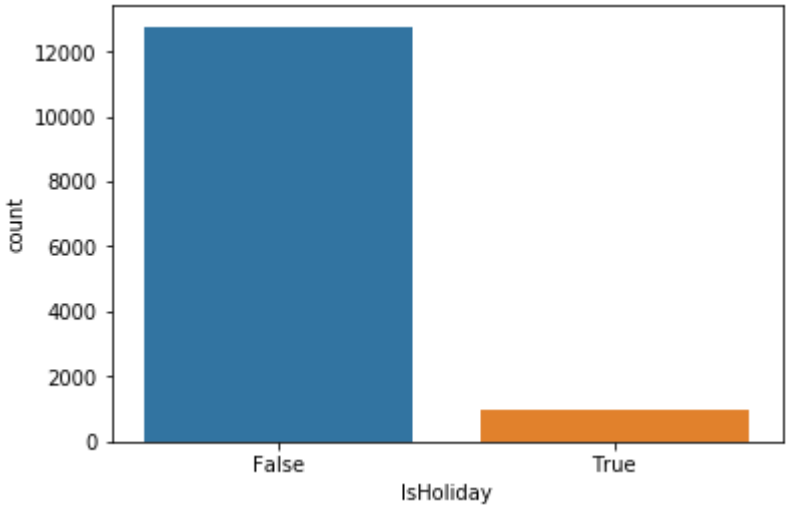

Fig 4.3 : Count plot of IsHoliday

4.2 Comparison of the Machine Learning Algorithms

The algorithms are compared by Accuracy, Mean Absolute Error and Mean Squared Error.

4.2.1 Comparison of Accuracy

Accuracy is the measure of prediction that our model got right. Accuracy is calculated by dividing number of correct prediction by total number of prediction.

Accuracy comes out to 0.88 or 88% means that the number of correct predictions are 88 out of 100 total examples.

The table below shows the accuracy after implementing each of the algorithm.

Table 4.3 : Comparison of accuracy of the algorithm

Algorithm	**Accuracy for training dataset**	**Accuracy for test dataset**
Linear Regression	86.94%	85.91%
Decision Tree	100%	83.71%
Decision Tree(max_depth=3)	86.94%	85.91%
Random Forest	98.25%	90.29%

The figure below shows a comparison of the accuracy of the algorithms.

Fig 4.4 : Comparison of Accuracy

4.2.2 Comparison of MAE and MSE

Mean absolute error (MAE) is the result of measuring the difference between two continuous variables.

Mean squared error (MSE) measures the average of squares of the errors that is the average squared difference between the estimated value and what is estimated.

The table below shows a comparison of mean squared error and mean absolute error after implementing each of the algorithms.

Table 4.4 : Comparison of MAE and MSE of the algorithm

Algorithm	Mean Absolute Error	Mean Squared Error
Linear Regression	3593.9453	58350291.2395
Decision Tree	3748.4436	67436571.3659
Decision Tree(max_depth=3)	4123.5649	68851663.8786
Random Forest	2956.3558	40201810.3052

The figures below shows the comparison of MAE and MSE.

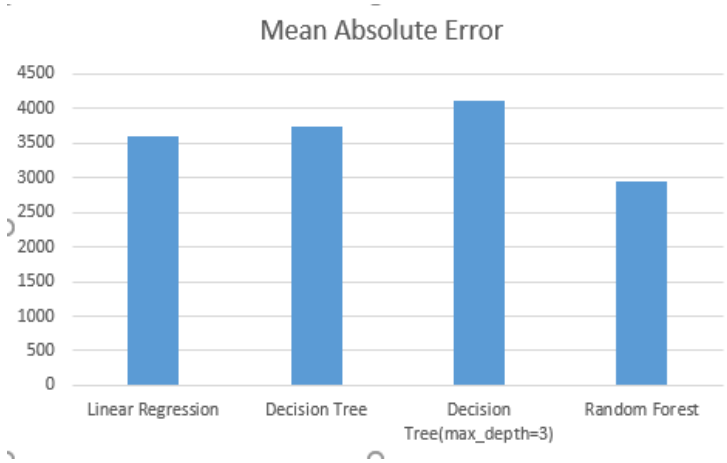

Fig 4.5 : Comparison of MAE

Fig 4.6 : Comparison of MSE

4.3 Result of Sales Forecasting System

We have predicted successfully weekly sales for each of the department of all the six stores. Sample of predicted weekly sales of some departments of store one are shown in the table below :

Table 4.5 : Sample of predicted weekly sales

Store Number	Department Number	Predicted Weekly Sales
1	1	24924.50
1	2	46039.49
1	3	41595.55
1	4	19403.54
1	5	21827.90
1	6	21043.39
1	7	22136.64
1	8	26229.21
1	9	57258.43
1	10	42960.91

From the comparative study among machine learning algorithms, we can decide which algorithm fits best. The best model has maximum accuracy and minimum Mean Absolute Error (MAE) and Mean Squared Error (MSE). Maximum accuracy of training data and test data is found by fitting the dataset into this algorithm. The value of MAE and MSE is also the smallest on Random Forest Regression. So, after comparing all the three results, we can say that random forest is the best algorithm for sales forecasting.

Chapter 5

CONCLUSION

With the gradually increasing number of business organizations, Sales forecasting has become very essential for any organization for having a successful future which is the basis of sales budget and production budget. It helps in deciding policies and facilitates in deciding the extent of advertising etc. Sales forecasting helps in preparing production and purchasing schedules. For the purpose of decision making, accurate sales forecasting is very important. Sales forecasting helps management to control supply chain. By anticipating future sales, organizations can make decisions about hiring permanent or temporarily.

5.1 Limitations of our system

For our project, we have only predicted weekly sales by using the dataset provided by kaggle. We have only discussed about a few algorithms but there are many other algorithms which can give better outcome. Real world data is usually never perfect. It is often noisy and have missing values. Data mining research have produced several methods of dealing with such data, including interpolation of missing values, binning, clustering etc. In the context of neural network based sales forecasting system, research has established that removing training tuples with missing values is the wisest approach.

5.2 Future Work

We have plan that we will do our project by using more real world data in future. More algorithms can be included to test the dataset. We will try to build our system totally based on artificial intelligence. The goal of most of the business organization is continuous improvement. By forecasting sales and continually revising the process to improve the accuracy, we can improve all aspects of business performances.

Chapter 6

REFERENCES

1. CHARLOTTE BOURNE.(2016, December 7). *Forecasting with Machine Learning Techniques* retrieved by February 16, 2019 from https://www.cardinalpath.com/forecasting-with-machine-learning-techniques.
2. Machine Learning (n.d.) In Wikipedia. Retrieved on February 16, 2019 from https://en.wikipedia.org/wiki/Machine_learning.
3. SUSAN WARD (November 22, 2018). *Sales Forecasting for New Businesses is Harder But Still Necessary.* Retrieved on February 16, 2019 from https://www.thebalancesmb.com/sales-forecasting-2948317.
4. SHREYASI GHOSE (n.d.). *Sales Forecasting: Meaning, Factors, Importance and Limitations.* Retrieved on February 16, 2019 from http://www.yourarticlelibrary.com/sales/sales-forecasting-meaning-factors-importance-and-limitations/50997
5. *Sales Force Composite Method.* Retrieved on February 16, 2019 from https://businessjargons.com/sales-force-composite-method.html
6. *Jury Method.* Retrieved on February 17, 2019 from https://businessjargons.com/jury-method.html.
7. SIDDHARTH KALLA (Apr 10, 2009). *Statistical Treatment of Data.* Retrieved Feb 17, 2019 from Explorable.com : https://explorable.com/statistical-treatment-of-data
8. MARKETING MANAGEMENT TUTORIAL. *SHORT, MEDIUM AND LONG TERM FORECASTING MARKETING MANAGEMENT*. Retrieved Feb 17, 2019 from https://www.wisdomjobs.com/e-university/marketing-management-tutorial-294/short-medium-and-long-term-forecasting-9586.html

9. NIKHILA C (n. d.) . *Sales Forecasting: Meaning, Importance and Methods*. Retrieved Feb 17, 2019 from http://www.businessmanagementideas.com/sales/forecasting-sales/sales-forecasting-meaning-importance-and-methods/7122
10. MITCHEL, T.M. 1997. *Machine Learning*. McGraw-Hill, New York, NY.
11. STAN MACK. *Long-term vs Short-term Forecasting for the Apparel Forecasting Process.* Retrieved Feb 17, 2019 from https://smallbusiness.chron.com/longterm-vs-shortterm-forecasting-apparel-forecasting-process-35753.html
12. JUSTIN DONAGHY on Quora (January 30, 2019). *What are the advantages and disadvantages of Machine Learning?* Retrieved Feb 18, 2019 from https://www.quora.com/What-are-the-advantages-and-disadvantages-of-machine-learning.
13. A. A. LEVIS AND L. G. PAPAGEORGIOU. *"Customer Demand Forecasting via Support Vector Regression Analysis".* In: Chemical Engineering Research and Design 8 (2005), pp. 1009–1018.
14. GIANNI DI PILLO et al. *An application of learning machines to sales forecasting under promotions.* Tech. rep. Sapienza, Universiteta di Roma, 2013.
15. MAIKE KRAUSE-TRAUDES ET AL. *Spatial data mining for retail sales forecasting. Tech. rep. Fraunhofer-Institute Intelligente Analyse- und Information systems (IAIS),* 2008.
16. *Drugs store sales forecast using Machine Learning* by HONGYU XIONG (hxiong2), Xi Wu (wuxi), Jingying Yue (jingying)
17. WENSEN DAI ET AL. "*A Clustering-based Sales Forecasting Scheme Using Support Vector Regression for Computer Server.*" Procedia Manufacturing 2 (2015) 82 – 86.
18. MARCO HULSMANN ET AL. *"General Sales Forecast Models for Automobile Markets and their Analysis." Transactions on Machine Learning and Data Mining* Vol. 5, No. 2 (2012) 65-86.
19. BERKOVEC, J.*: New Car Sales and Used Car Stocks: A Model for the Automobile Market*, The RAND Journal of Economics 2, 195–214 (1985)
20. OUN ABBAS, KHURRAM SHAHZAD, GHULAM ALI, UMAR SARWAR on *Issues on Sales Forecasting for Apparel Industry.*

21. ALEJANDRO BALDOMINOS, IVÁN BLANCO, ANTONIO JOSÉ MORENO, RUBÉN ITURRARTE, CARLOS AFONSO (21 November, 2018) . *Identifying Real Estate Opportunities Using Machine Learning*
22. KAGGLE (2014). *Walmart Recruiting – Store Sales Forecasting*. Retrieved Feb 17, 2019 from https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data.

Chapter 7

APPENDIX

Appendix A : Abbreviations

DT : Decision Tree

RF: Random Forest

CIP : Consumer Price Index

MAE : Mean Absolute Error

MSE : Mean Squared Error

Appendix B : Source Code

```
# -*- coding: utf-8 -*-
"""
Created on Tue Jan  8 17:53:41 2019

@author: Subah
"""

import pandas as pd
ds = pd.read_csv('train.csv')

print("Shape of dataset:", ds.shape)
ds.dtypes

Data_dict = pd.DataFrame(ds.dtypes)
Data_dict['MissingVal'] = ds.isnull().sum()
Data_dict['UniqueVal'] = ds.nunique()
Data_dict['Count'] = ds.count()
Data_dict = Data_dict.rename(columns = {0:'DataType'})
print("Print dictionary details: \n",Data_dict)
```

```
import seaborn as sns
sns.countplot(ds['Store'],label="Count")
sns.countplot(ds['Dept'],label="Count")
sns.countplot(ds['IsHoliday'],label="Count")

ds['Date'] = pd.to_datetime(ds['Date'])
ds.head()

df = pd.get_dummies(ds, columns=['Store', 'Dept'])
df.head()

df['Date_dayofweek'] = df['Date'].dt.dayofweek
df['Date_month'] = df['Date'].dt.month
df['Date_year'] = df['Date'].dt.year
df['Date_day'] = df['Date'].dt.day

for days_to_lag in [1, 2, 3, 5, 7, 14, 30]:
    df['Weekly_sales_lag_{}'.format(days_to_lag)] =
df.Weekly_Sales.shift(days_to_lag)

df.head()

df = df.fillna(0)
```

```
df.IsHoliday = df.IsHoliday.astype(int)

x = df[df.columns.difference(['Date', 'Weekly_Sales'])]
y = df.Weekly_Sales
x.head()
y[:5]

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=1/3)
print("Shape of training dataset:",x_train.shape)
print("Shape of testing dataset:",x_test.shape)

#Linear Regression
from sklearn.linear_model import LinearRegression
clf = LinearRegression()
clf.fit(x_train, y_train)

y_pred = clf.predict(x_test)

print("Accuracy on training dataset using Linear Regreession:
{:.4f}".format(clf.score(x_train, y_train)))
print("Accuracy on test dataset using Linear Regreession:
{:.4f}".format(clf.score(x_test, y_test)))
```

```
#Decision Tree Regressor
from sklearn.tree import DecisionTreeRegressor
regressor = DecisionTreeRegressor(random_state = 0)
regressor.fit(x_train, y_train)
y_pred1 = regressor.predict(x_test)
print("Accuracy on training dataset using Decision Tree Regreession:
{:.4f}".format(regressor.score(x_train, y_train)))
print("Accuracy on test dataset using Decision Tree Regreession:
{:.4f}".format(regressor.score(x_test, y_test)))

regressor = DecisionTreeRegressor(max_depth = 3, random_state = 0)
regressor.fit(x_train, y_train)
y_pred2 = regressor.predict(x_test)

print("Accuracy on training dataset using Decision Tree Regreession(max_depth =
3): {:.4f}".format(clf.score(x_train, y_train)))
print("Accuracy on test dataset using Decision Tree Regreession(max_depth = 3):
{:.4f}".format(clf.score(x_test, y_test)))

#Random Forest Regressor
from sklearn.ensemble import RandomForestRegressor
clf = RandomForestRegressor()
clf.fit(x_train, y_train)
y_pred3 = clf.predict(x_test)
```

```
print("Accuracy on training dataset using Random Forest :
{:.4f}".format(clf.score(x_train, y_train)))

print("Accuracy on test dataset using Random Forest:
{:.4f}".format(clf.score(x_test, y_test)))


from sklearn.metrics import mean_squared_error, mean_absolute_error


#for Linear Regression

print("MAE of Linear Regreession: {:.4f}".format(mean_absolute_error(y_test,
y_pred)))

print("MSE of Linear Regreession: {:.4f}".format(mean_squared_error(y_test,
y_pred)))


#for Decision Tree regression

print("MAE of Decision Tree Regreession:

{:.4f}".format(mean_absolute_error(y_test, y_pred1)))

print("MSE of Decision Tree Regreession:
{:.4f}".format(mean_squared_error(y_test, y_pred1)))


#for Decision Tree Regression(max_depth = 3)

print("MAE of Decision Tree Regression(max_depth = 3):
{:.4f}".format(mean_absolute_error(y_test, y_pred2)))

print("MSE of Decision Tree Regression(max_depth = 3):
{:.4f}".format(mean_squared_error(y_test, y_pred2)))
```

```
#for Random Forest
print("MAE of Random Forest: {:.4f}".format(mean_absolute_error(y_test, y_pred3)))
print("MSE of Random Forest: {:.4f}".format(mean_squared_error(y_test, y_pred3)))
```