SALES PREDICTION USING MACHINE LEARNING

Minor Project Report

Submitted in partial fulfilment of the
requirements for the award of the degree

Of

Bachelor of Technology
In

Computer Science and Engineering

BY

Ms. Somya Sharma(Roll No.-8817103032)

Mr. Animesh Kumar Jain(Roll No.-8817103008)

Under the guidance of

Dr. Nishant Srivastava

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JAYPEE UNIVERSITY ANOOPSHAHR – 223390, UP (INDIA)

MAY, 2020

JAYPEE UNIVERSITY ANOOPSHAHR

CERTIFICATE

This is to certify that the report titled **"Sales Prediction Using Machine Learning"** submitted by **Ms. Somya Sharma** and **Mr. Animesh Kumar Jain** in partial fulfilment of the requirements for the award of Bachelor of Technology degree in **Computer Science and Engineering** during session 2019-2020 at **Japyee University Anoopshahr** is an authentic work by them under my supervision and guidance.

To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other university / institute for the award of any Degree.

Date:______________________

Dr. Nishant Srivastava

Head of Department

Dept. of Computer Science and Engineering

Japyee University Anoopshahr

____

ACKNOWLEDGEMENT

We would like to express our gratitude to our mentor **Dr. Nishant Srivastava** for his guidance, advice and constant support throughout our project work. We would like to thank him for being our advisor.

Next, we want to express our respects to **Mr. Lalit Mohan Gupta and Dr. Mohammad Qasim Rafiq** for teaching us and also helping us how to learn. They have been great sources of inspiration to us and we thank them from the bottom of my heart.

We would like to thank all our friends and especially our classmates for all the thoughtful and mind stimulating discussions we had, which prompted us to think beyond the obvious. We are especially indebted to our parents for their love, sacrifice, and support and would like to thank our parents for raising us in a way to believe that we can achieve anything in life with hard work and dedication.

Ms. Somya Sharma (Roll Number: - 8817103032)

Mr. Animesh Kumar Jain (Roll Number: - 8817103008)

INDEX

CONTENTS **PAGE**

ABSTRACT

The ability to predict data accurately is extremely valuable in a vast array of domains such as stocks, sales, weather or even sports. Presented here is the study and implementation of several algorithms employed on sales data, consisting of weekly retail sales numbers from different departments in Walmart retail outlets all over the United States of America. The models implemented for prediction are Linear Regression, Decision Tree and Random Forest. The hyperparameters of each model were varied to obtain the best value of R^2 score. The project aims to incorporate state-of-the-art technique for predicting sales with the goal of achieving high accuracy. In this project, we use a completely machine learning based approach to solve the problem of accurate prediction. The machine is trained on the dataset Walmart Store Sales Forecasting. The resulting system is fast and accurate, thus aiding those applications and organizations which require sales prediction.

1 INTRODUCTION

1.1 PROBLEM STATEMENT

Many household products are sold by various subsidiaries of the retail store network which are geographically located at various locations. Supply chain inefficiencies will occur at different locations when the market potential will not evaluated by the retailers. Many times it is not easy for the retailers to understand the market condition at various geographical locations. The organization of retail store network has to understand the market conditions to intensify its goods to be bought and sold so that many number of customers get attracted in that direction. Business forecast helps retailers to visualize the big picture by forecasting the sales we get a general idea of coming years if any changes are needed then those changes are done in the retail store's objective so that success is achieved more profitably .It also helps the customers to be happy by providing the products desired by them in desired time, when the customers are happy then they prefer the store that provides all the resources they need to their satisfaction by this the sales in the particular store in which the customers purchase more items increases causing more profit. The forecasting of sales helps to know the retailers the demand of the product. Therefore, the goal is to implement a model which has a great accuracy in predicting weekly sales.

1.2 OBJECTIVE

We aim to learn about the machine learning techniques for regression models that are used to predict numeric data like sales and to implement a system for better accuracy in that prediction.

1.3 MACHINE LEARNING PARADIGM

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. In other words, Machine Learning can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training our machines (computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate. Learning algorithms can be of three types.

1.3.1 SUPERVISED LEARNING

This algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Ours is a supervised learning model.

1.3.2 UNSUPERVISED LEARNING

In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.

1.3.3 REINFORCEMENT LEARNING

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

1.4 OVERVIEW

Our sales prediction system can be divided into the following steps:-

1.4.1 DATA PRE-PROCESSING

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. For example, some algorithms like Random Forest does take a dataset containing null values. Thus, all the null values need to be removed or replaced with some other values.

1.4.2 FEATURE ENGINEERING

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by selecting relevant or creating features from data that help facilitate the machine learning process.

1.4.3 MODEL TRAINING AND TESTING

Then using the processed data different machine learning algorithms are trained and tested to make predictions. Prediction algorithm can be a classification or regression one. Classification algorithm is the one where output is a discrete class while Regression algorithm is the one where output is a continuous quantity. Here, we have used different regression models like Linear Regression, Decision Tree and Random Forest.

1.4.4 PREDICTION RESULT ASSESSMENT

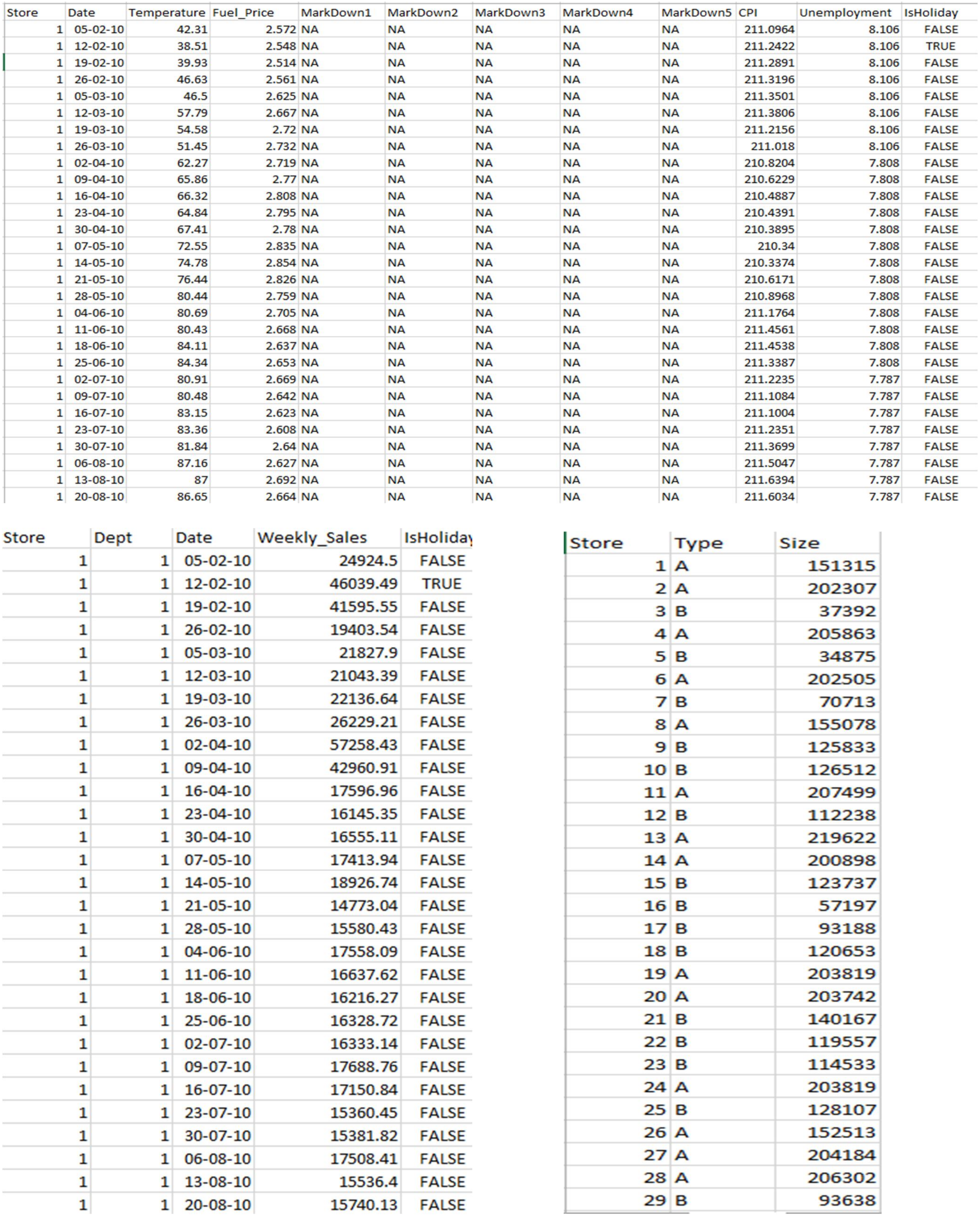
The goal of any machine learning project is to implement a model using an algorithm which would predict most accurately on the given dataset. For measuring the accuracy of a prediction algorithm we may use several metrics like Mean Square Error, Mean Absolute Error, R^2 Score and many more. We have used R^2 Score as an accuracy measure for our algorithms.

1.5 DATASET DESCRIPTION

In this project, we have downloaded a historic data of 45 Walmart stores located in different regions which cover the sales from 2010-02-05 to 2012-11-01. The dataset is spread into three different CSV file names features, store and train having common columns. The CSV files contain the following fields:

- Store - The store number
- Dept - The department number
- Date - The week
- Weekly_Sales - Sales for the given department in the given store
- IsHoliday - Whether the week is a special holiday week
- Temperature - Average temperature in the region
- Fuel_Price - Cost of fuel in the region
- MarkDown1-5 - Anonymized data related to promotional markdowns of Walmart

- CPI - The consumer price index
- Unemployment - The unemployment rate
- Size- Size of the store
- Type- Type of store



Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
1	05-02-10	42.31	2.572	NA	NA	NA	NA	NA	211.0964	8.106	FALSE
1	12-02-10	38.51	2.548	NA	NA	NA	NA	NA	211.2422	8.106	TRUE
1	19-02-10	39.93	2.514	NA	NA	NA	NA	NA	211.2891	8.106	FALSE
1	26-02-10	46.63	2.561	NA	NA	NA	NA	NA	211.3196	8.106	FALSE
1	05-03-10	46.5	2.625	NA	NA	NA	NA	NA	211.3501	8.106	FALSE
1	12-03-10	57.79	2.667	NA	NA	NA	NA	NA	211.3806	8.106	FALSE
1	19-03-10	54.58	2.72	NA	NA	NA	NA	NA	211.2156	8.106	FALSE
1	26-03-10	51.45	2.732	NA	NA	NA	NA	NA	211.018	8.106	FALSE
1	02-04-10	62.27	2.719	NA	NA	NA	NA	NA	210.8204	7.808	FALSE
1	09-04-10	65.86	2.77	NA	NA	NA	NA	NA	210.6229	7.808	FALSE
1	16-04-10	66.32	2.808	NA	NA	NA	NA	NA	210.4887	7.808	FALSE
1	23-04-10	64.84	2.795	NA	NA	NA	NA	NA	210.4391	7.808	FALSE
1	30-04-10	67.41	2.78	NA	NA	NA	NA	NA	210.3895	7.808	FALSE
1	07-05-10	72.55	2.835	NA	NA	NA	NA	NA	210.34	7.808	FALSE
1	14-05-10	74.78	2.854	NA	NA	NA	NA	NA	210.3374	7.808	FALSE
1	21-05-10	76.44	2.826	NA	NA	NA	NA	NA	210.6171	7.808	FALSE
1	28-05-10	80.44	2.759	NA	NA	NA	NA	NA	210.8968	7.808	FALSE
1	04-06-10	80.69	2.705	NA	NA	NA	NA	NA	211.1764	7.808	FALSE
1	11-06-10	80.43	2.668	NA	NA	NA	NA	NA	211.4561	7.808	FALSE
1	18-06-10	84.11	2.637	NA	NA	NA	NA	NA	211.4538	7.808	FALSE
1	25-06-10	84.34	2.653	NA	NA	NA	NA	NA	211.3387	7.808	FALSE
1	02-07-10	80.91	2.669	NA	NA	NA	NA	NA	211.2235	7.787	FALSE
1	09-07-10	80.48	2.642	NA	NA	NA	NA	NA	211.1084	7.787	FALSE
1	16-07-10	83.15	2.623	NA	NA	NA	NA	NA	211.1004	7.787	FALSE
1	23-07-10	83.36	2.608	NA	NA	NA	NA	NA	211.2351	7.787	FALSE
1	30-07-10	81.84	2.64	NA	NA	NA	NA	NA	211.3699	7.787	FALSE
1	06-08-10	87.16	2.627	NA	NA	NA	NA	NA	211.5047	7.787	FALSE
1	13-08-10	87	2.692	NA	NA	NA	NA	NA	211.6394	7.787	FALSE
1	20-08-10	86.65	2.664	NA	NA	NA	NA	NA	211.6034	7.787	FALSE

Store	Dept	Date	Weekly_Sales	IsHoliday
1	1	05-02-10	24924.5	FALSE
1	1	12-02-10	46039.49	TRUE
1	1	19-02-10	41595.55	FALSE
1	1	26-02-10	19403.54	FALSE
1	1	05-03-10	21827.9	FALSE
1	1	12-03-10	21043.39	FALSE
1	1	19-03-10	22136.64	FALSE
1	1	26-03-10	26229.21	FALSE
1	1	02-04-10	57258.43	FALSE
1	1	09-04-10	42960.91	FALSE
1	1	16-04-10	17596.96	FALSE
1	1	23-04-10	16145.35	FALSE
1	1	30-04-10	16555.11	FALSE
1	1	07-05-10	17413.94	FALSE
1	1	14-05-10	18926.74	FALSE
1	1	21-05-10	14773.04	FALSE
1	1	28-05-10	15580.43	FALSE
1	1	04-06-10	17558.09	FALSE
1	1	11-06-10	16637.62	FALSE
1	1	18-06-10	16216.27	FALSE
1	1	25-06-10	16328.72	FALSE
1	1	02-07-10	16333.14	FALSE
1	1	09-07-10	17688.76	FALSE
1	1	16-07-10	17150.84	FALSE
1	1	23-07-10	15360.45	FALSE
1	1	30-07-10	15381.82	FALSE
1	1	06-08-10	17508.41	FALSE
1	1	13-08-10	15536.4	FALSE
1	1	20-08-10	15740.13	FALSE

Store	Type	Size
1	A	151315
2	A	202307
3	B	37392
4	A	205863
5	B	34875
6	A	202505
7	B	70713
8	A	155078
9	B	125833
10	B	126512
11	A	207499
12	B	112238
13	A	219622
14	A	200898
15	B	123737
16	B	57197
17	B	93188
18	B	120653
19	A	203819
20	A	203742
21	B	140167
22	B	119557
23	B	114533
24	A	203819
25	B	128107
26	A	152513
27	A	204184
28	A	206302
29	B	93638

Fig: Dataset.

- This data set has been downloaded from

  https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

- In our dataset, we have total 15 features and 1 target value.

- In total there are 4,21,570 entries in the dataset.

2 TECHNIQUE USED

2.1 FOR DATA PRE-PROCESSING

For making of dataset ready so that it may give higher accuracy on the prediction algorithms we have used the following techniques:

2.1.1 IMPUTATION

Missing values are one of the most common problems WE can encounter when WE try to prepare your data for machine learning. The reason for the missing values might be human errors, interruptions in the data flow, privacy concerns, and so on. Whatever is the reason, missing values affect the performance of the machine learning models. The simplest solution to the missing values is to drop the rows or the entire column. But that leads to loss of data. So, Imputation is a more preferable option rather than dropping because it preserves the data size. Now, the most important thing to keep in mind was to have a sensible value that will replace all the missing value. In our dataset, all the null values were in the MarkDown fields (which is effect of promotional schemes run by Walmart of the weekly sales). Thus, it can be 0. So, all the null (NA) values were replaced with 0.

2.1.2 CONVERSION OF DATA

Some of the prediction algorithms work only on numerical data resulting into the need for converting our categorical data into numeric data. In our dataset, we found that IsHoliday and Type were of boolean and character data type respectively. Thus they were converted into numeric form. For boolean data type 0 was taken as False and 1 as True. For character data type we took A as 1, B as 2 and C as 3. Thus the whole dataset was converted into numeric form.

2.1.3 EXTRACTING DATE

Though date columns usually provide valuable information about the model target, they are neglected as an input or used nonsensically for the machine learning algorithms. The reason behind this is that dates can be present in numerous formats, which make it hard to be understood by algorithms. Thus, creating a new and simplified format is essential from the date feature to have an effect on the predictions. In our dataset, we have extracted the parts of the date into separate columns: Year, Month and Day.

2.2 FOR FEATURE ENGINEERING

Not all the features in the dataset always have an effect on the target prediction value. Thus, it becomes an important task to get rid of those features that are least or not significant. To evaluate this relationship between the feature and target we have used Pearson's Correlation. After evaluation we found IsHoliday, Temperature and Fuel Prize were the least effecting the weekly sales. Thus, they were dropped from the data. .

2.2.1 PEARSON'S CORRELATION

The Pearson correlation coefficient (named for Karl Pearson) can be used to summarize the strength of the linear relationship between two data samples.

The Pearson's correlation coefficient is calculated as the covariance of the two variables divided by the product of the standard deviation of each data sample. It is the normalization of the covariance between the two variables to give an interpretable score.

The formula for Pearson's correlation coefficient is given by:

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Where, r_{XY} is the coefficient, $\bar{X}$ and $\bar{Y}$ are the mean of the data samples.

The result of the calculation, the correlation coefficient can be interpreted to understand the relationship.

The coefficient returns a value between -1 and 1 that represents the limits of correlation from a full negative correlation to a full positive correlation. A value of 0 means no correlation. The value must be interpreted, where often a value below -0.5 or above 0.5 indicates a notable correlation, and values below those values suggests a less notable correlation.

The Pearson's correlation coefficient can be used to evaluate the relationship between more than two variables.

This can be done by calculating a matrix of the relationships between each pair of variables in the dataset. The result is a symmetric matrix called a correlation matrix with a value of 1.0 along the diagonal as each column always perfectly correlates with itself.

To visualize this relationship of the target value with all the features we have used heatmaps. A heatmap is a pictorial representation of the values using color coding.

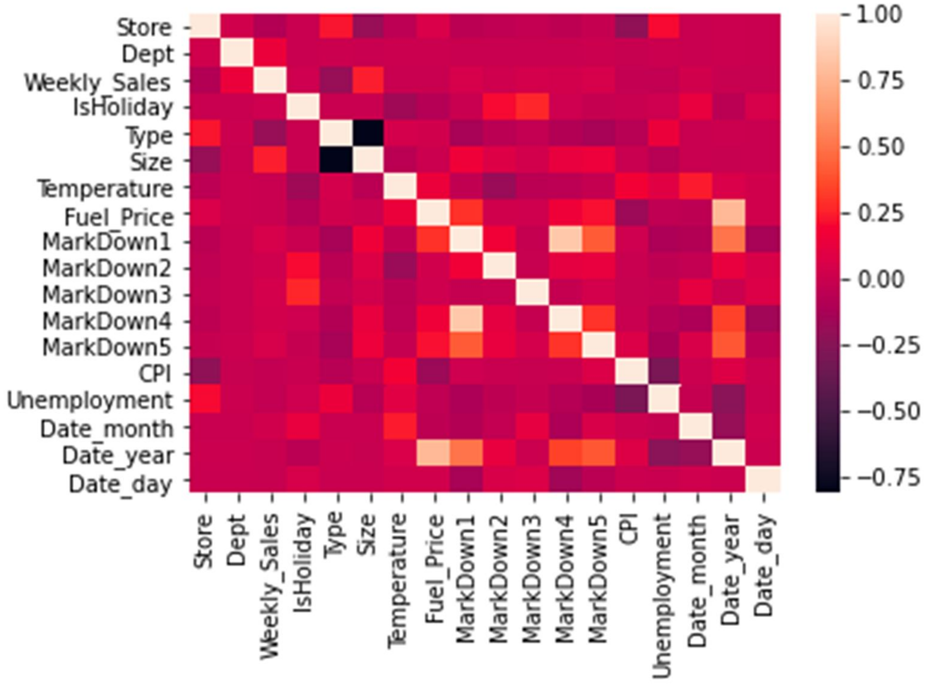


Fig. Sample of a Pearson's Correlation Matrix Heatmap.

2.3 FOR MODEL TRAINING AND TESTING

Our model is based on regression and we have taken three different regression algorithms for prediction.

2.3.1 LINEAR REGRESSION

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. Instead of simple linear regression, where we have one predictor and one outcome, we will go with multiple linear regression, where we have more than one predictors and one outcome. Linear regression looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the points to the regression line. In case of multi linear regression this line becomes a multi-dimensional plane. Multi Linear Regression follows the formula:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

$$\textbf{where, for } i = n \textbf{ observations:}$$

$$y_i = \text{dependent variable}$$

$$x_i = \text{expanatory variables}$$

$$\beta_0 = \text{y-intercept (constant term)}$$

$$\beta_p = \text{slope coefficients for each explanatory variable}$$

$$\epsilon = \text{the model's error term (also known as the residuals)}$$

The coefficients in this linear equation denote the magnitude of additive relation between the predictor and the response. In simpler words, keeping everything else fixed, a unit change in x1 will lead to change of β1 in the outcome, and so on. Thus while calculating a coefficient we will neglect other terms in the equation.

So, the following step are to be taken to implement multi linear regression.

STEP 1: We will calculate the mean value for the dependent and all the independent variables.

STEP 2: Then using the following formula we will calculate the coefficient.

$$\beta_i = \frac{mean(y)}{mean(X_i)}$$

Where, β_i is the coefficient of i th independent value (X_i) and y is the dependent variable.

STEP 3: After calculating all the coefficients we will put them in the equation along with mean value of all the variables and calculate the y-intercept (β_0). Thus we will have the equation for the best fit regression plane or line,

STEP 4: Now, we will plot all the points of dependent and independent variables on a graphical plane. Along with the regression plane made from the equation we have calculated.

STEP 5: Now by extrapolating the line or plane we can predict the target value.

The distance between the actual and predicted values is the error (ϵ). We may reduce the error by using different planes for regression by using the hit and trial method for n number of planes.

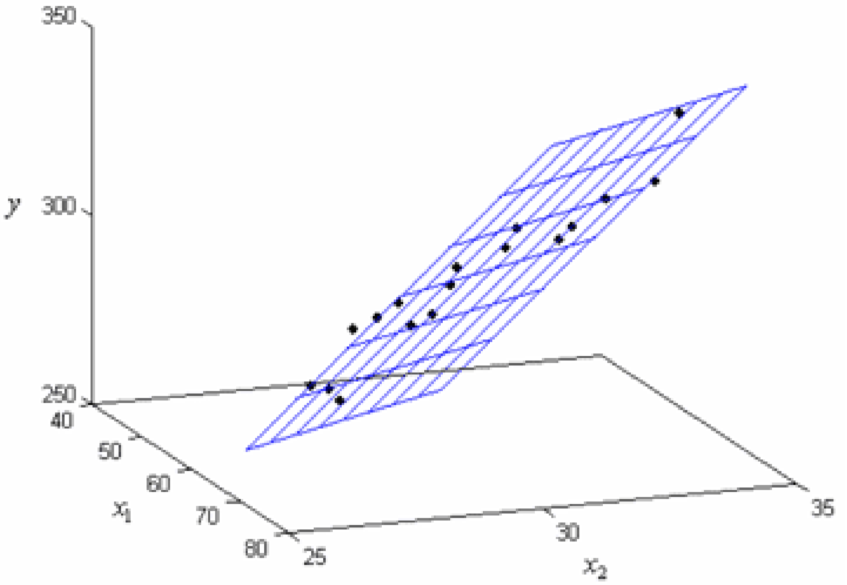


Fig. Sample of a regression plane.

2.3.2 DECISION TREE

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain (used for classification) with Standard Deviation Reduction.

STANDARD DEVIATION

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous). We use standard deviation to calculate the homogeneity of a numerical sample. If the numerical sample is completely homogeneous its standard deviation is zero.

a) Standard deviation for one attribute is given by:

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Where, x is the observation, $\bar{x}$ is the mean of observation and n is the total number of observations.

b) Standard deviation for two attribute is given by:

$$S(T,X) = \sum_{c \in X} P(c)S(c)$$

Where $P(c)$ is the probability of occurrence of c and $S(c)$ is the standard deviation of c.

STANDARD DEVIATION REDUCTION

The standard deviation reduction is based on the decrease in standard deviation after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest standard deviation reduction (i.e., the most homogeneous branches).

The following steps are taken to construct the decision tree:

STEP 1: The standard deviation of the target is calculated.

STEP 2: The dataset is then split on the different attributes. The standard deviation for each branch is calculated. The resulting standard deviation is subtracted from the standard deviation before the split. The result is the standard deviation reduction.

STEP 3: The attribute with the largest standard deviation reduction is chosen for the decision node.

STEP 4: The dataset is divided based on the values of the selected attribute. This process is run recursively on the non-leaf branches, until all data is processed.

In practice, we need some termination criteria. For that when coefficient of deviation (CV) for a branch becomes smaller than a certain threshold and/or when too few instances (n) remain in the branch. Coefficient of deviation is calculated using the formula:

$$CV = \frac{Standard\ Deviation\ of\ x}{Mean\ of\ x} = \frac{S}{\bar{x}}$$

Note, when the number of instances is more than one at a leaf node we calculate the average as the final value for the target.

Thus, we have the decision tree. Now for every set of input we see the leaf node it leads to and that is the predicted value.

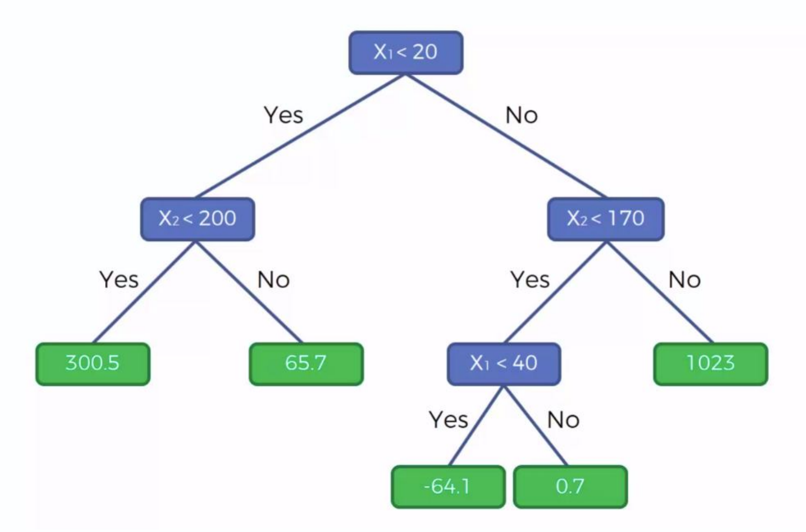


Fig. Sample of a decision tree

2.3.3 RANDOM FOREST

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression**.** An Ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.

Random forest uses the Bootstrap Aggregation, commonly known as bagging. It refers to random sampling with replacement. Bootstrap allows us to better understand the bias and the variance with the dataset. Bootstrap involves random sampling of small subset of data from the dataset.

It is a general procedure that can be used to reduce the variance for those algorithm that have high variance, like decision trees. Bagging makes each model run independently and then aggregates the outputs at the end without preference to any model**.**

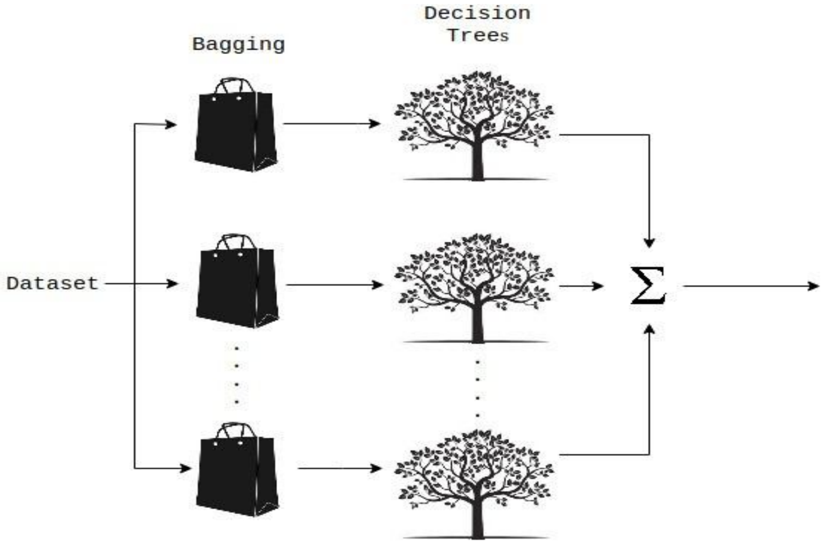


Fig. Bagging

Random forest algorithm can be divided into the following steps:

STEP 1: Pick at random some data points from the dataset, these will be our sampled training datasets.

STEP 2: Now, train a decision tree for each sampled training dataset.

STEP 3: Repeat the above steps 'n' times i.e. the no. of decision trees we want to build. Each time random samples with replacement are selected for training.

STEP 4: Predict the target value using the testing data for each decision tree.

STEP 5: Now assign the final predicted value as the average of all the predicted value of each decision tree.

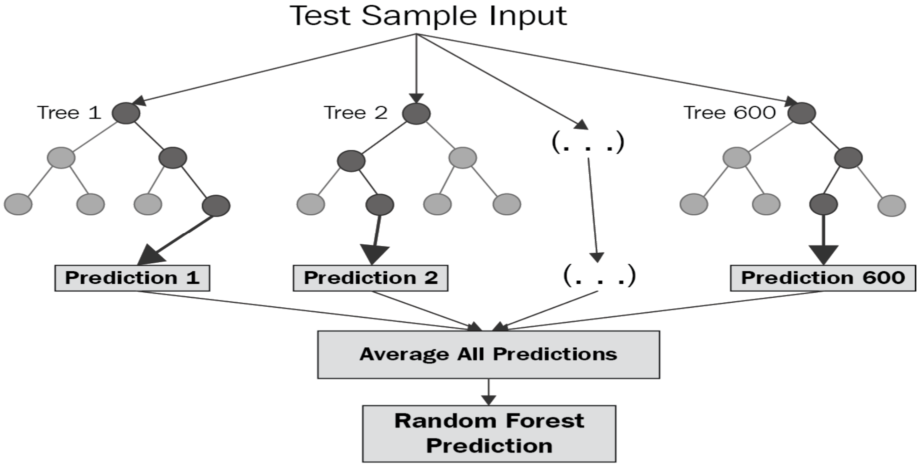


Fig. Structure of Random Forest

2.4 FOR PREDICTION RESULT ASSESSMENT

After the model for prediction is build. An important task is to check how accurately the model is predicting on the given test data. For regression models we have several metrics like Mean Squared Error, Mean Absolute Error, R Squared Error and many more. In our project we have taken R Squared Error metrics as a measure to check accuracy.

2.4.1 R SQUARED ERROR

R squared represents the proportion of variance that has been explained by the independent variables in the model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, through the proportion of explained variance. It is also known as the coefficient of determination.

The formula for R Squared is as follows:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where, $\hat{y}i$ is the predicted value for ith sample, y is the actual value for ith sample, $\bar{y}i$ is the mean value of the data

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). Although in our project we have converted the score into percentage to see what percentage accuracy does the model have.

3 EXPERIMENTAL RESULT

3.1 IMPLEMENTATION DETAILS

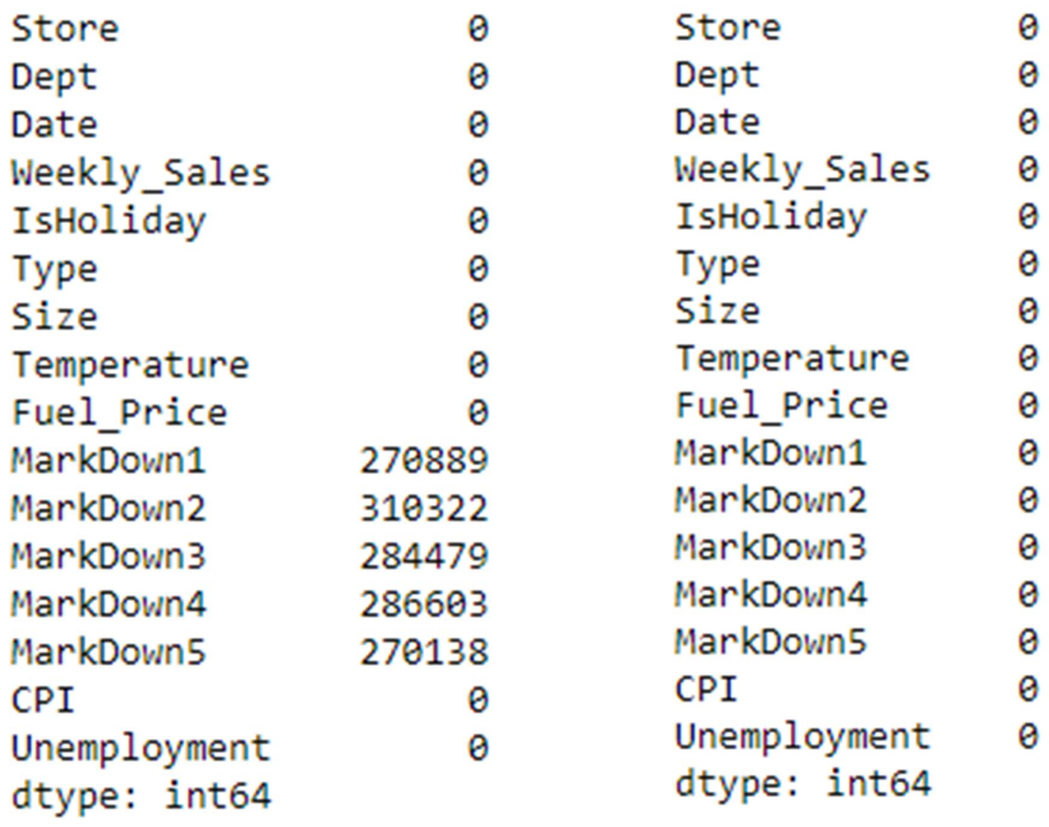
The project is implemented in python 3. Python libraries like SKLearn along with Numpy, Pandas and Seaborn was used for implementation. The system specifications on which the model is trained and evaluated are mentioned as follows:

CPU - Intel Core i5- 3320M 2.60 GHz

RAM – 8 GB DDR3.

3.2 PROCESSED DATA

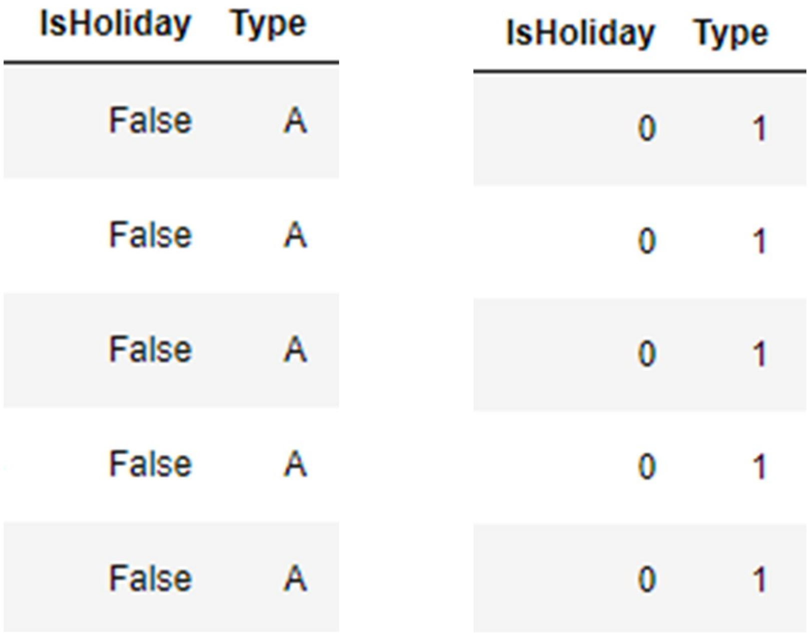
These are the results of data processing:



```
Store                0
Dept                 0
Date                 0
Weekly_Sales         0
IsHoliday            0
Type                 0
Size                 0
Temperature          0
Fuel_Price           0
MarkDown1       270889
MarkDown2       310322
MarkDown3       284479
MarkDown4       286603
MarkDown5       270138
CPI                  0
Unemployment         0
dtype: int64
```

```
Store           0
Dept            0
Date            0
Weekly_Sales    0
IsHoliday       0
Type            0
Size            0
Temperature     0
Fuel_Price      0
MarkDown1       0
MarkDown2       0
MarkDown3       0
MarkDown4       0
MarkDown5       0
CPI             0
Unemployment    0
dtype: int64
```

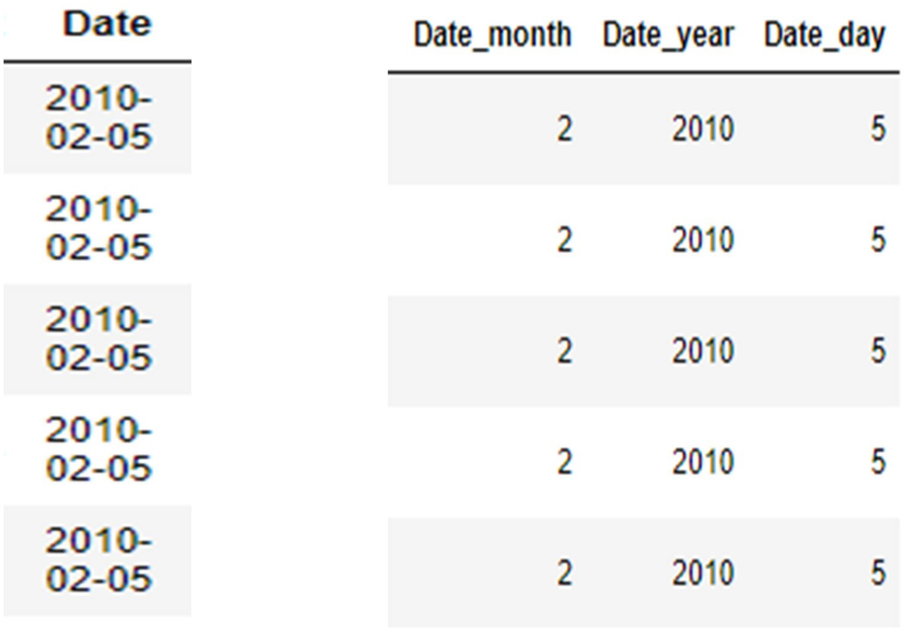
Fig. Total Null values before and after processing.



IsHoliday	Type
False	A
False	A
False	A
False	A
False	A

IsHoliday	Type
0	1
0	1
0	1
0	1
0	1

Fig. Data before and after type conversion (from categorical to numeric).



Date
2010-02-05
2010-02-05
2010-02-05
2010-02-05
2010-02-05

Date_month	Date_year	Date_day
2	2010	5
2	2010	5
2	2010	5
2	2010	5
2	2010	5

Fig. Date before and after separating into different columns.

3.3 CORRELATION MATRIX

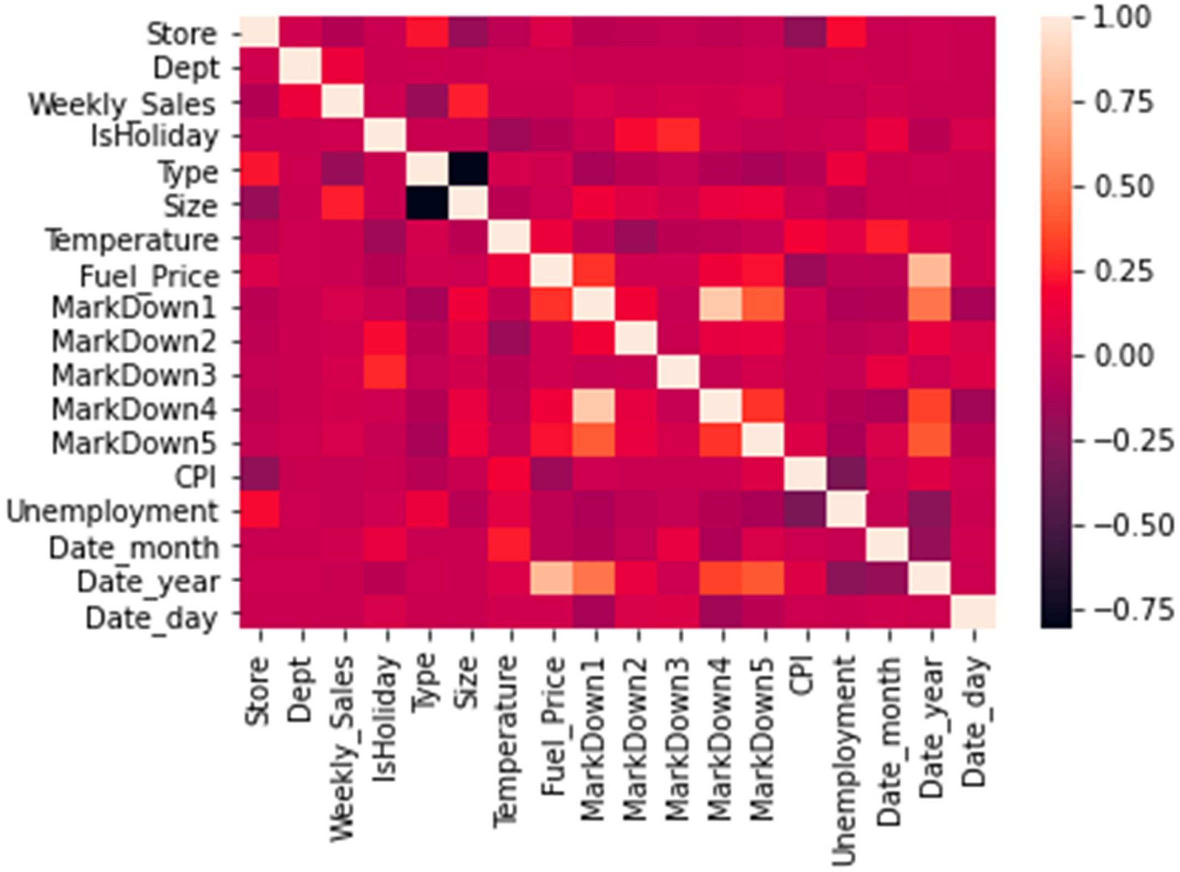
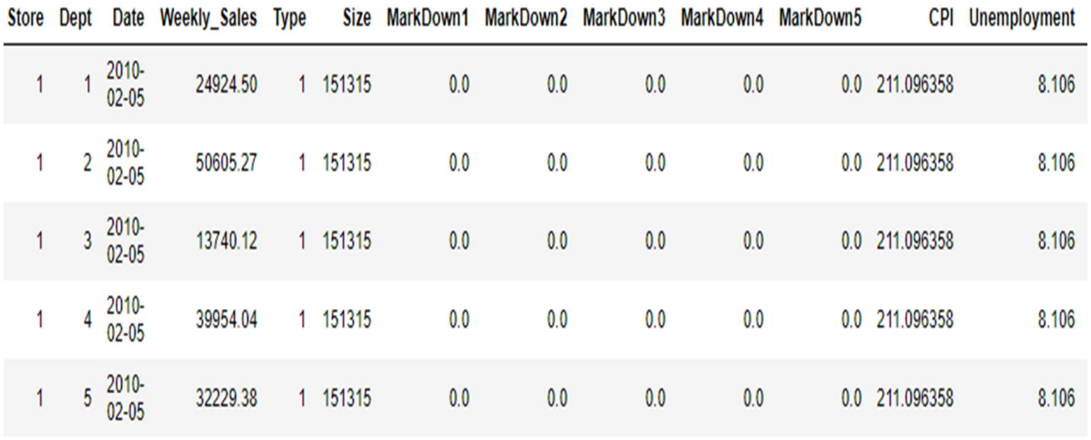
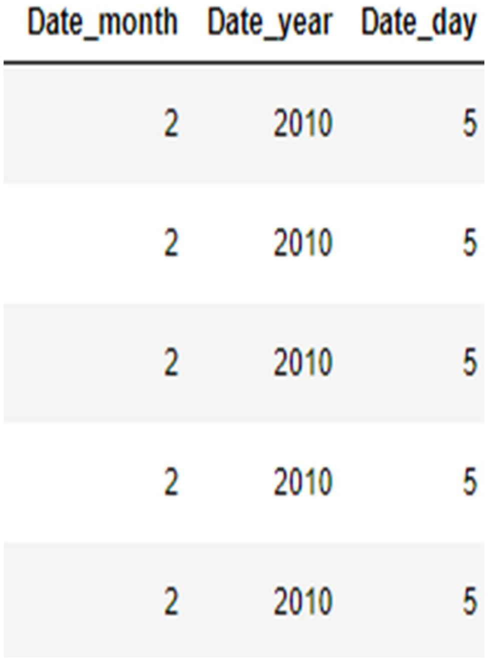


Fig. Pearson’s correlation matrix’s heatmap.



Store	Dept	Date	Weekly_Sales	Type	Size	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
1	1	2010-02-05	24924.50	1	151315	0.0	0.0	0.0	0.0	0.0	211.096358	8.106
1	2	2010-02-05	50605.27	1	151315	0.0	0.0	0.0	0.0	0.0	211.096358	8.106
1	3	2010-02-05	13740.12	1	151315	0.0	0.0	0.0	0.0	0.0	211.096358	8.106
1	4	2010-02-05	39954.04	1	151315	0.0	0.0	0.0	0.0	0.0	211.096358	8.106
1	5	2010-02-05	32229.38	1	151315	0.0	0.0	0.0	0.0	0.0	211.096358	8.106

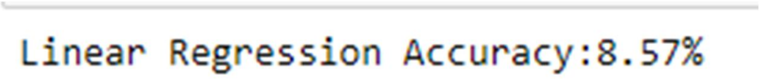
Fig. Final data for training and testing.



Date_month	Date_year	Date_day
2	2010	5
2	2010	5
2	2010	5
2	2010	5
2	2010	5

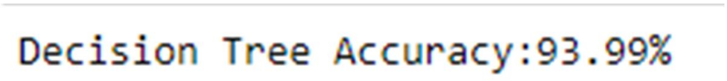
Fig. Final data for training and testing.

3.4 PREDICTION RESULTS



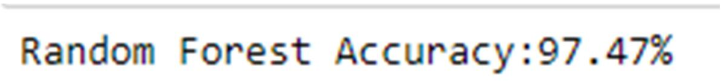
```
Linear Regression Accuracy:8.57%
```

Fig: Prediction accuracy using linear regression.



```
Decision Tree Accuracy:93.99%
```

Fig: Prediction accuracy using decision tree.



```
Random Forest Accuracy:97.47%
```

Fig. Prediction accuracy using random forest.

4 LIMITATIONS

Our prediction model will not give such accuracy if training and testing is done with 50% of the dataset. And computation time will increase for comparatively larger datasets. Our system won't work is any of the data fields are missing. And it won't work for anyone type of data. Also it takes into account data of 2 years only thus won't be accurate for predicting values after 4-5 years as the market trends will be changed by then.

5 FUTURE ENHANCEMENTS

Other approaches with different algorithms like SVM, Neural Networks, ARIMA and many more can be done for better accuracy over comparatively larger datasets.

6 CONCLUSION

An accurate and efficient Sales Prediction model has been developed which achieves comparable metrics with the existing state-of-the-art system. This project uses recent techniques in the field of regression analysis and machine learning. Already studied and famous dataset of Walmart Sales was used and the evaluation was consistent. This can be used in real-time applications which require sales prediction. After training and testing we got 97.47% accuracy using random forest for our model.

7 REFERENCES:

- Anita S. Harsoor, Anushree Patil, **“FORECAST OF SALES OF WALMART STORE USING BIG DATA APPLICATIONS”** in International Journal of Research in Engineering and Technology, Volume 4, Issue 6, June 2015.

- Dataset-
  https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

- Article-
  https://medium.com/@aravanshad/how-to-choose-machine-learning-algorithms-9a92a448e0df
  https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114
  https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e
  https://saedsayad.com/decision_tree_reg.htm

- Documentation of python libraries-
  https://scikit-learn.org/stable/modules/model_evaluation.html

- Video Lectures-
  https://www.youtube.com/watch?v=g9c66TUylZ4
  https://www.youtube.com/watch?v=J4Wdy0Wc_xQ
  https://www.youtube.com/watch?v=E5RjzSK0fvY