Capstone Project

Rossmann Sales Prediction

Team Members

Arslan Ullah Khan
Jayanti Mala

Points for Discussion

1. Introduction
2. Problem statement
3. Data Summary
4. Approach
5. Sales with respect to years
6. Impact of promo on sales
7. Sales with respect to assortment
8. Sales with respect to school holiday
9. Sales on different days of week
10. Sales with respect to store type
11. Sales outliers detection and removal
12. Checking correlation
13. Linear Regression Model
14. Decision Tree Regression
15. Random Forest Regression
16. Important Features
17. Conclusion

Introduction:

An important part of present-day business intelligence is sales prediction. Sales prediction can be termed a complex problem, and it gets harder in the case of lack of data or missing data values, and the presence of outliers.

This dataset is a live dataset of Roseman Stores. On analysing this problem we observe that Roseman problem is a regression problem and our primarily goal is to predict the sales figures of Roseman problem. Sales forecasting plays an integral role in setting expectations and making plans for your business. It’s your best shot at predicting the future.

Problem Statement

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

We are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column.

Data Summary

1. Id - an Id that represents a (Store, Date) duple within the test set
2. Store - a unique Id for each store
3. Sales - the turnover for any given day (this is what you are predicting)
4. Customers - the number of customers on a given day
5. Open - an indicator for whether the store was open: 0 = closed, 1 = open
6. StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
7. SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools

8. StoreType - differentiates between 4 different store models: a, b, c, d

9. Assortment - describes an assortment level: a = basic, b = extra, c = extended

10. CompetitionDistance - distance in meters to the nearest competitor store

11. CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened

12. Promo - indicates whether a store is running a promo on that day

13. Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating

14. Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2

15. PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

Approach

1. Importing the required libraries and reading the dataset.

   a. Merging of the two datasets

   b. Understanding the dataset

2. Exploratory Data Analysis (EDA) –

   a. Data Visualization

3. Feature Engineering

   a. Dropping of unwanted columns and values (closed stores)

   b. Filling Missing Values with Imputation

   c. Outliers Detection and removal

4. Label Encoding (Converting categorical variables to numerical values)

6. Model Building

    a. Linear Regression Model
    b. Decision Tree Regression Model
    c. Random Forest Regression Model

7. Model Validation

    a. r2 score
    b. Mean absolute error
    c. Root mean squared error

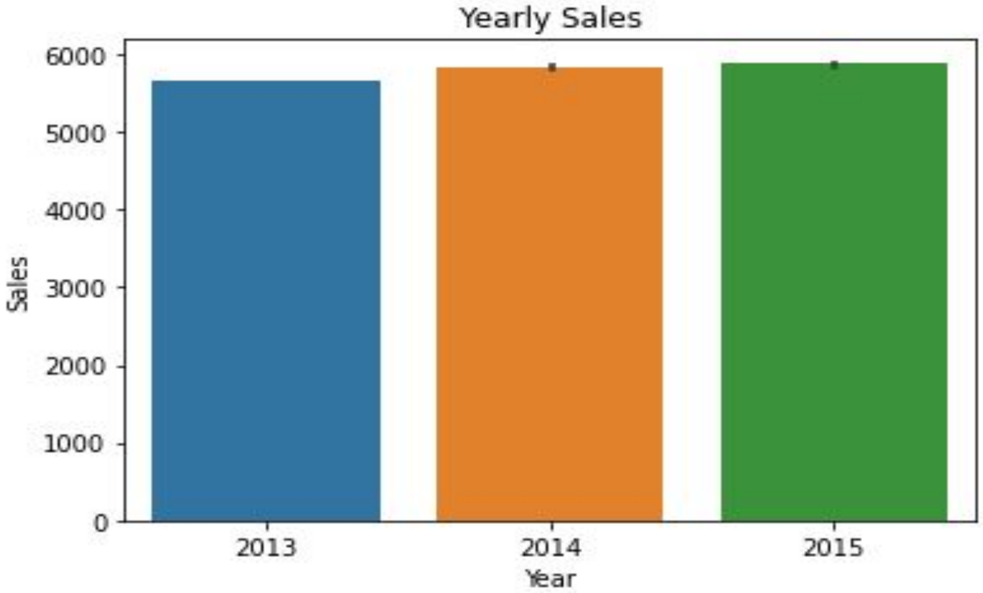
8. Creating the final right model and making predictions

9. Conclusion

Relation between different Variable

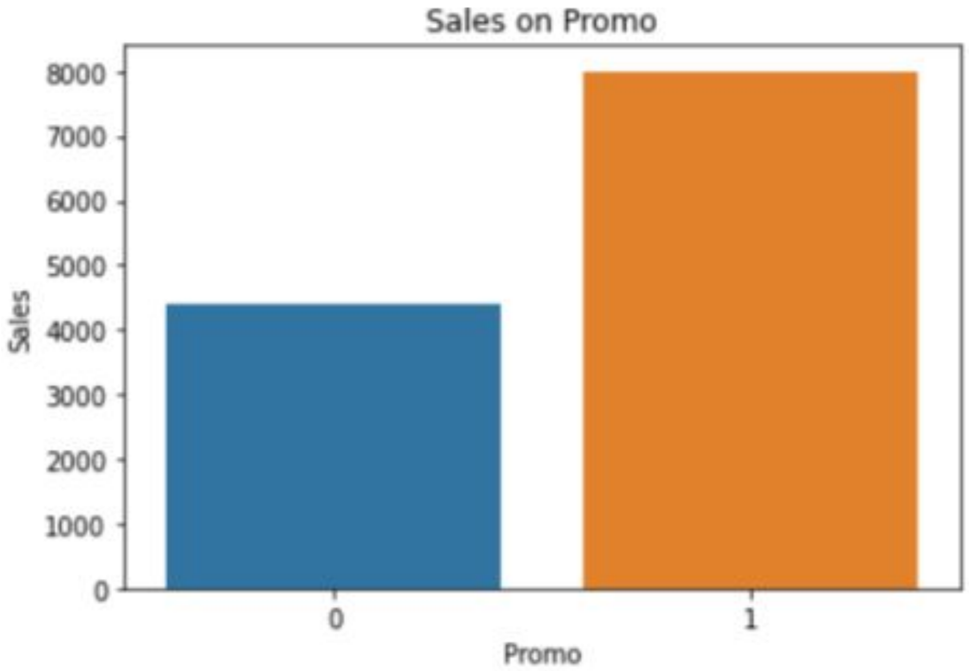
A. Sales with respect to year

Observation- Increasing in sales year to year



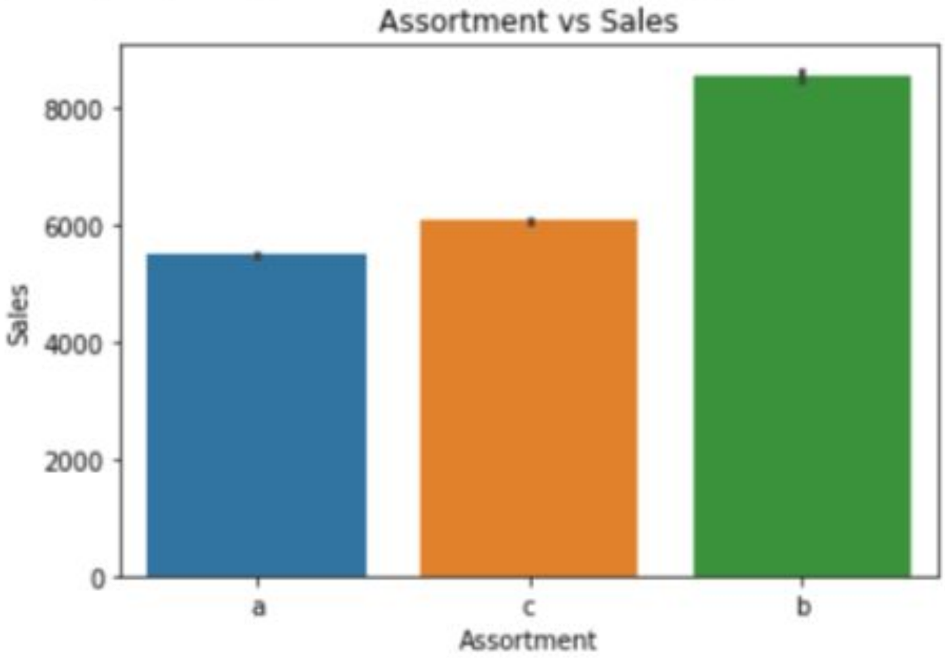
B. Impact of promos on sales

Observation - When promo are running we can expect more sales.



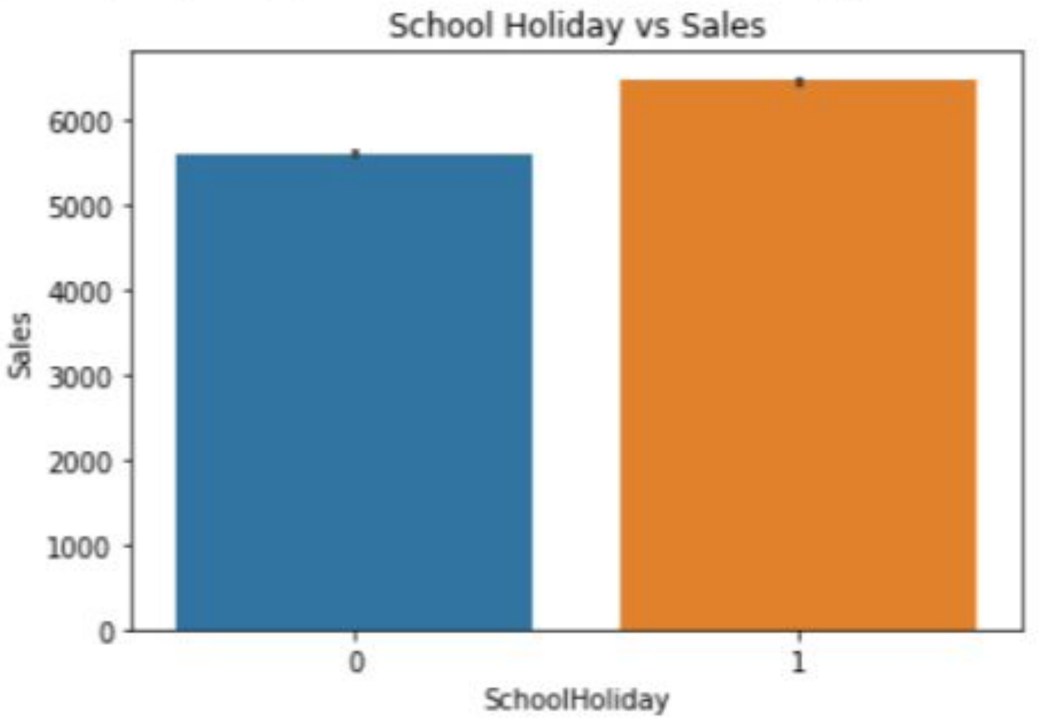
C. Sales with respect to Assortment

Observation - Assortment level 'b' have the highest sales



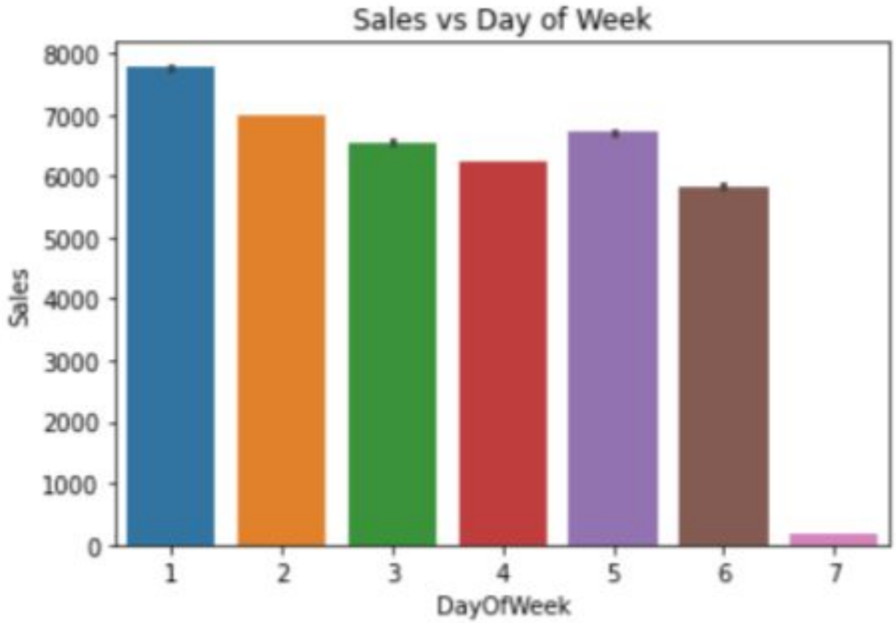
D. Sales with respect to School Holiday

Observation - More sales on School Holidays



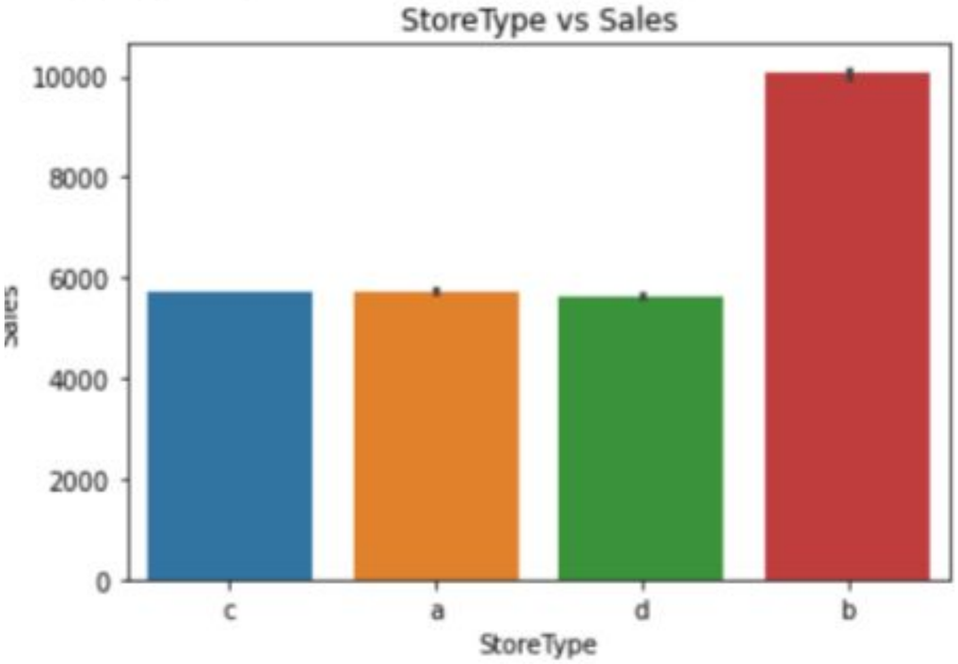
E. Sales on different days of the week

Observation - Monday, Tuesday and Friday have the highest sales and lowest sale on Sunday



F. Sales with respect to store type

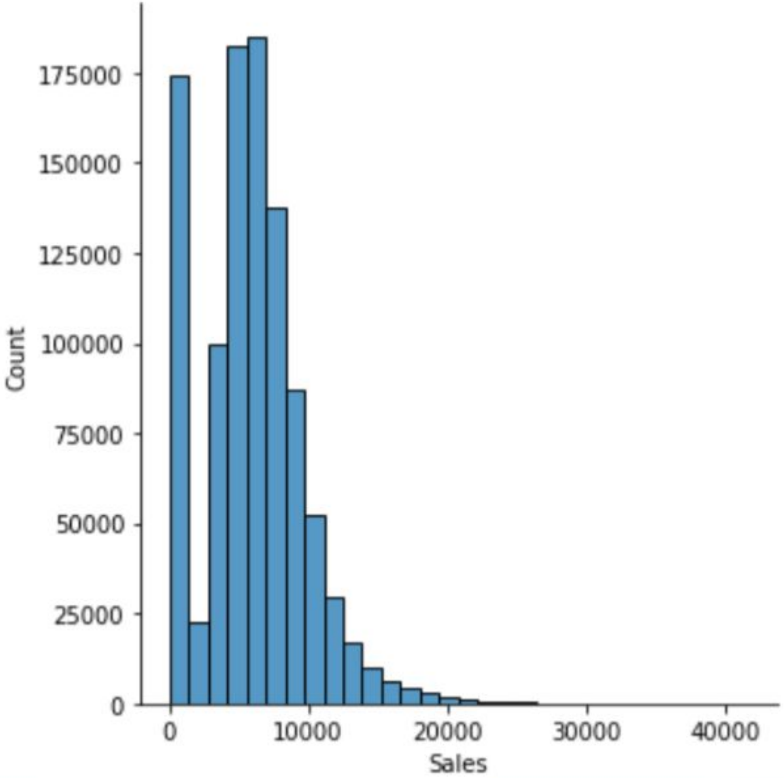
Observation - b type stores have the highest sales.



Sales Outliers Detection and Removal

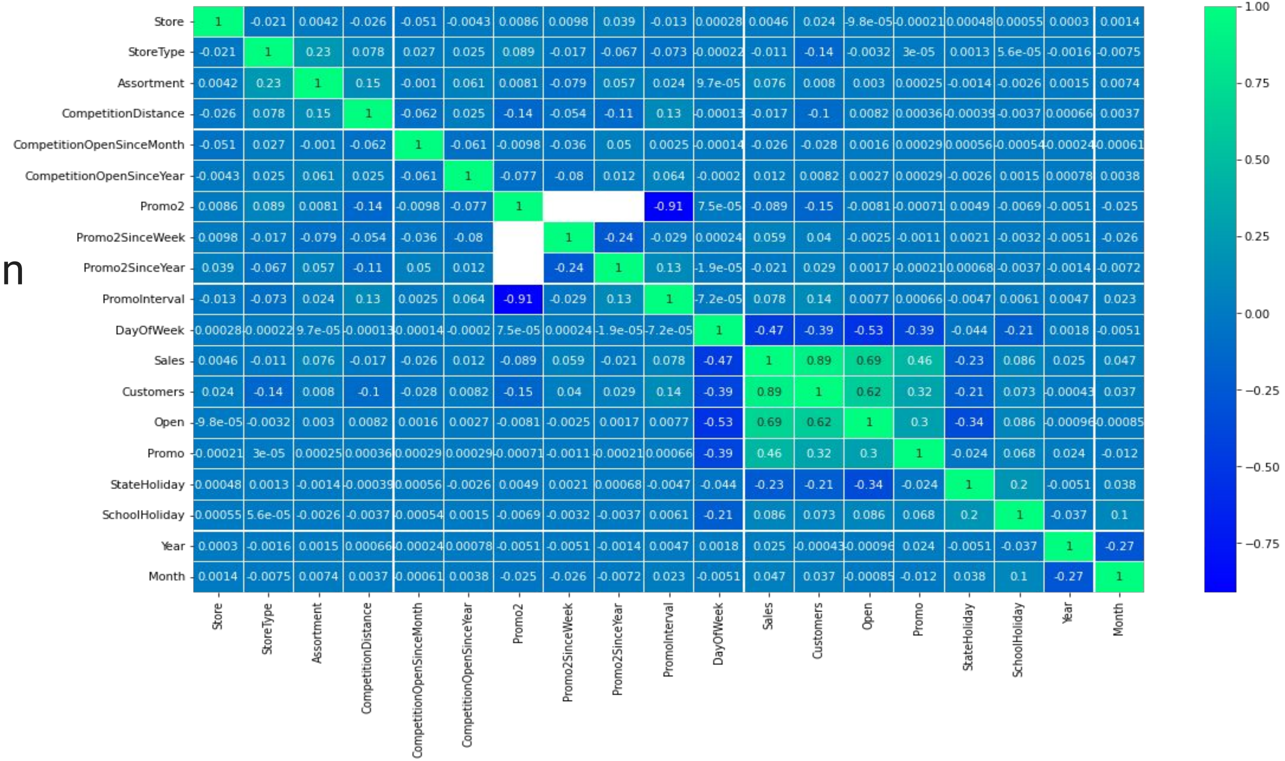
.

Observation - From the above graph we can see that sales >25k are very less, so it might be an outlier so we drop them.



Checking Correlation

Observation- Sales are highly correlated with Customers, Open and Promo code, and minorly correlated to school holidays

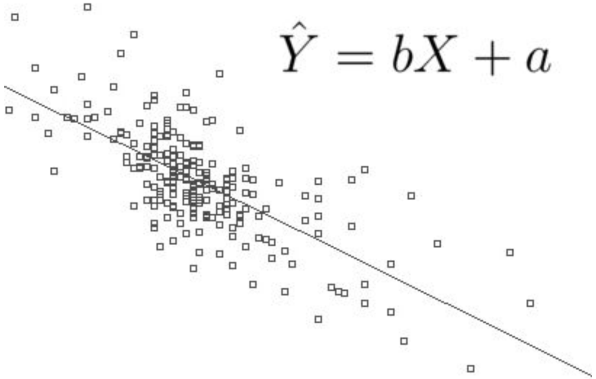


	Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	Year	Month
Store	1	-0.021	0.0042	-0.026	-0.051	-0.0043	0.0086	0.0098	0.039	-0.013	0.00028	0.0046	0.024	-9.8e-05	-0.00021	0.00048	0.00055	0.0003	0.0014
StoreType	-0.021	1	0.23	0.078	0.027	0.025	0.089	-0.017	-0.067	-0.073	-0.00022	-0.011	-0.14	-0.0032	3e-05	0.0013	5.6e-05	-0.0016	-0.0075
Assortment	0.0042	0.23	1	0.15	-0.001	0.061	0.0081	-0.079	0.057	0.024	9.7e-05	0.076	0.008	0.003	0.00025	-0.0014	-0.0026	0.0015	0.0074
CompetitionDistance	-0.026	0.078	0.15	1	-0.062	0.025	-0.14	-0.054	-0.11	0.13	-0.00013	-0.017	-0.1	0.0082	0.00036	-0.00039	-0.0037	0.00066	0.0037
CompetitionOpenSinceMonth	-0.051	0.027	-0.001	-0.062	1	-0.061	-0.0098	-0.036	0.05	0.0025	-0.00014	-0.026	-0.028	0.0016	0.00029	0.00056	-0.00054	-0.00024	-0.00061
CompetitionOpenSinceYear	-0.0043	0.025	0.061	0.025	-0.061	1	-0.077	-0.08	0.012	0.064	-0.0002	0.012	0.0082	0.0027	0.00029	-0.0026	0.0015	0.00078	0.0038
Promo2	0.0086	0.089	0.0081	-0.14	-0.0098	-0.077	1			-0.91	7.5e-05	-0.089	-0.15	-0.0081	-0.00071	0.0049	-0.0069	-0.0051	-0.025
Promo2SinceWeek	0.0098	-0.017	-0.079	-0.054	-0.036	-0.08		1	-0.24	-0.029	0.00024	0.059	0.04	-0.0025	-0.0011	0.0021	-0.0032	-0.0051	-0.026
Promo2SinceYear	0.039	-0.067	0.057	-0.11	0.05	0.012		-0.24	1	0.13	-1.9e-05	-0.021	0.029	0.0017	-0.00021	0.00068	-0.0037	-0.0014	-0.0072
PromoInterval	-0.013	-0.073	0.024	0.13	0.0025	0.064	-0.91	-0.029	0.13	1	-7.2e-05	0.078	0.14	0.0077	0.00066	-0.0047	0.0061	0.0047	0.023
DayOfWeek	0.00028	-0.00022	9.7e-05	-0.00013	-0.00014	-0.0002	7.5e-05	0.00024	-1.9e-05	-7.2e-05	1	-0.47	-0.39	-0.53	-0.39	-0.044	-0.21	0.0018	-0.0051
Sales	0.0046	-0.011	0.076	-0.017	-0.026	0.012	-0.089	0.059	-0.021	0.078	-0.47	1	0.89	0.69	0.46	-0.23	0.086	0.025	0.047
Customers	0.024	-0.14	0.008	-0.1	-0.028	0.0082	-0.15	0.04	0.029	0.14	-0.39	0.89	1	0.62	0.32	-0.21	0.073	-0.00043	0.037
Open	-9.8e-05	-0.0032	0.003	0.0082	0.0016	0.0027	-0.0081	-0.0025	0.0017	0.0077	-0.53	0.69	0.62	1	0.3	-0.34	0.086	-0.00096	-0.00085
Promo	-0.00021	3e-05	0.00025	0.00036	0.00029	0.00029	-0.00071	-0.0011	-0.00021	0.00066	-0.39	0.46	0.32	0.3	1	-0.024	0.068	0.024	-0.012
StateHoliday	0.00048	0.0013	-0.0014	-0.00039	0.00056	-0.0026	0.0049	0.0021	0.00068	-0.0047	-0.044	-0.23	-0.21	-0.34	-0.024	1	0.2	-0.0051	0.038
SchoolHoliday	0.00055	5.6e-05	-0.0026	-0.0037	-0.00054	0.0015	-0.0069	-0.0032	-0.0037	0.0061	-0.21	0.086	0.073	0.086	0.068	0.2	1	-0.037	0.1
Year	0.0003	-0.0016	0.0015	0.00066	-0.00024	0.00078	-0.0051	-0.0051	-0.0014	0.0047	0.0018	0.025	-0.00043	-0.00096	0.024	-0.0051	-0.037	1	-0.27
Month	0.0014	-0.0075	0.0074	0.0037	-0.00061	0.0038	-0.025	-0.026	-0.0072	0.023	-0.0051	0.047	0.037	-0.00085	-0.012	0.038	0.1	-0.27	1

Model Building

A. Linear Regression Model

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.



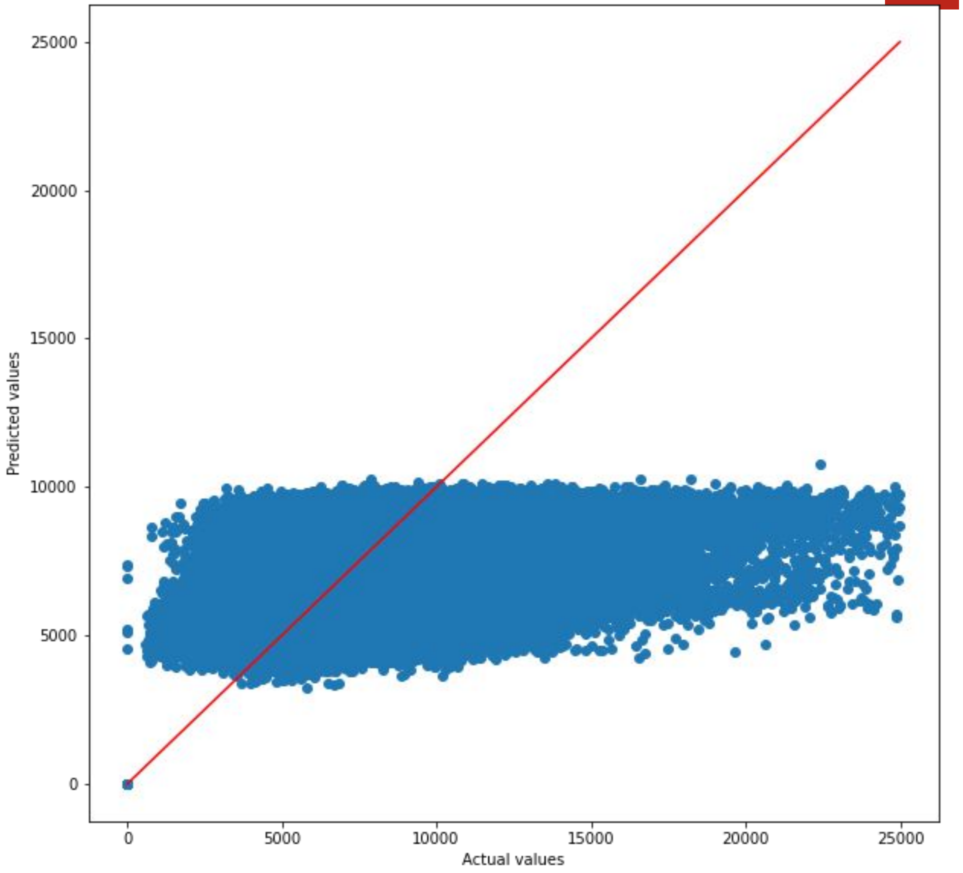
Performance of Model

- The r2_score of Linear Regression is: **0.7784061010317642**

- The Mean absolute error of Linear Regression is: **992.30**

- The Root mean squared error of Linear Regression is: **1914.395102807405**

Observation - From the above plot we can see that Linear regression model is performing badly as its not making any predictions more than 10000.



B. Decision Tree Regression

Decision Tree Regression:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

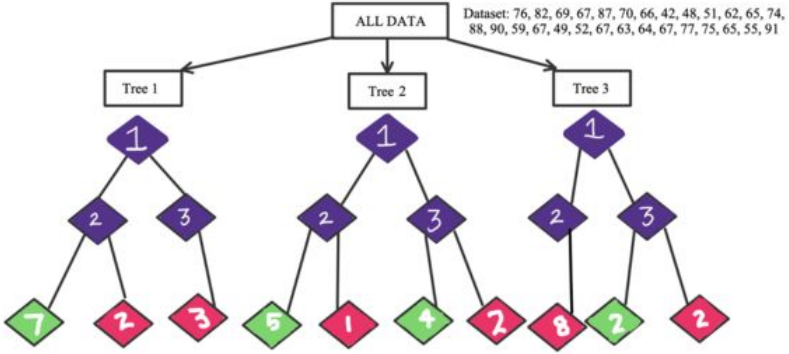
Discrete output example: A weather prediction model that predicts whether or not there'll be rain on a particular day.

Continuous output example: A profit prediction model that states the probable profit that can be generated from the sale of a product.

Performance of Model

- The r2_score = **0.9519176919198888**
- The Mean absolute error = **420.01**
- The Root mean squared error = **891.7545389502772**

Observation - The decision tree regressor performing well as compared to linear regressors


ALL DATA
Dataset: 76, 82, 69, 67, 87, 70, 66, 42, 48, 51, 62, 65, 74, 88, 90, 59, 67, 49, 52, 67, 63, 64, 67, 77, 75, 65, 55, 91
Tree 1
Tree 2
Tree 3
1
1
1
2
3
2
3
2
3
7
2
3
5
1
4
2
8
2
2

C. Random Forest Regressor

Random Forest Regression is a supervised learning algorithm that uses **ensemble learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

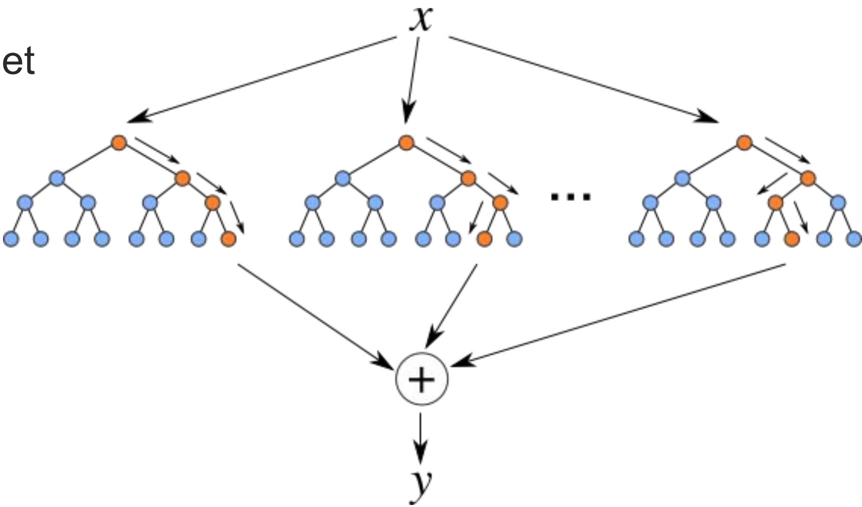
Step 1: Identify your dependent (y) and independent variables (X)

Step 2: Split the dataset into the Training set and Test set

Step 3: Training the Random Forest Regression

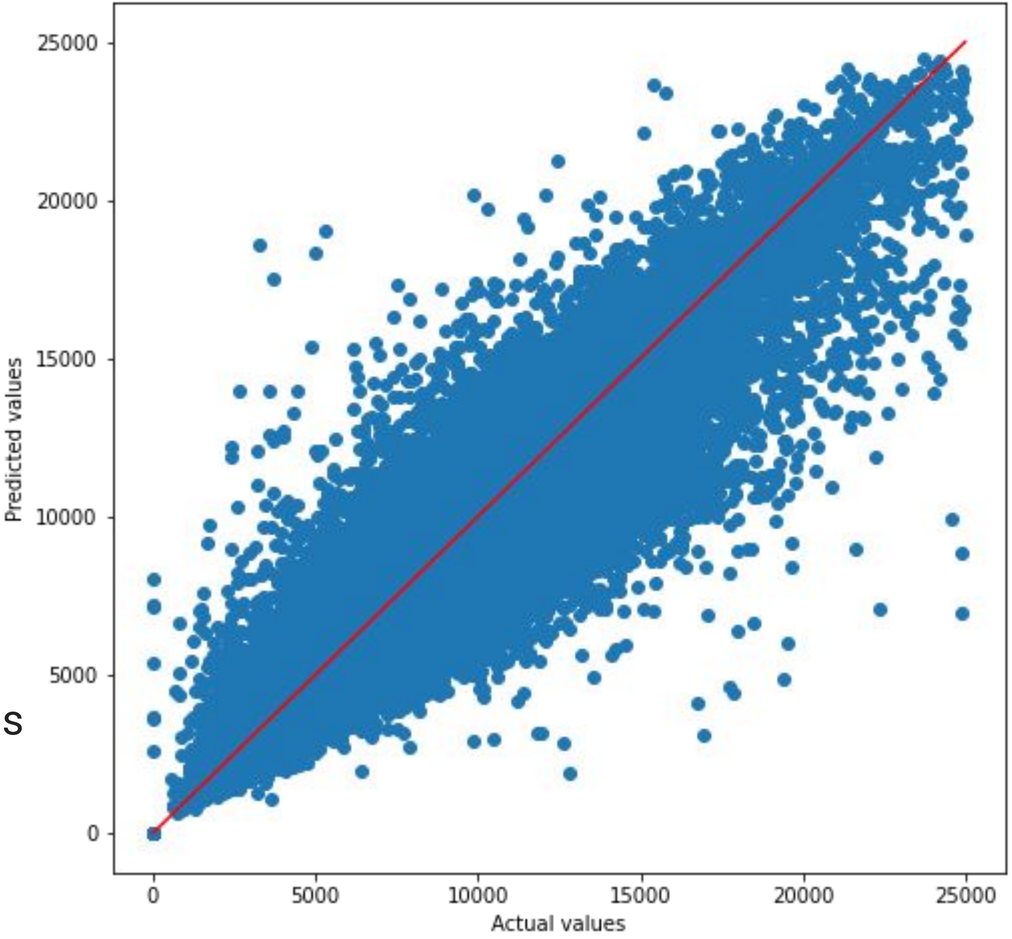
model on the whole dataset

Step 4: Predicting the Test set results



Performance of Model

- The r2_score: **0.9656108742155831**
- Mean absolute error: **355.77**
- Root mean squared error: **754.1595639323232**

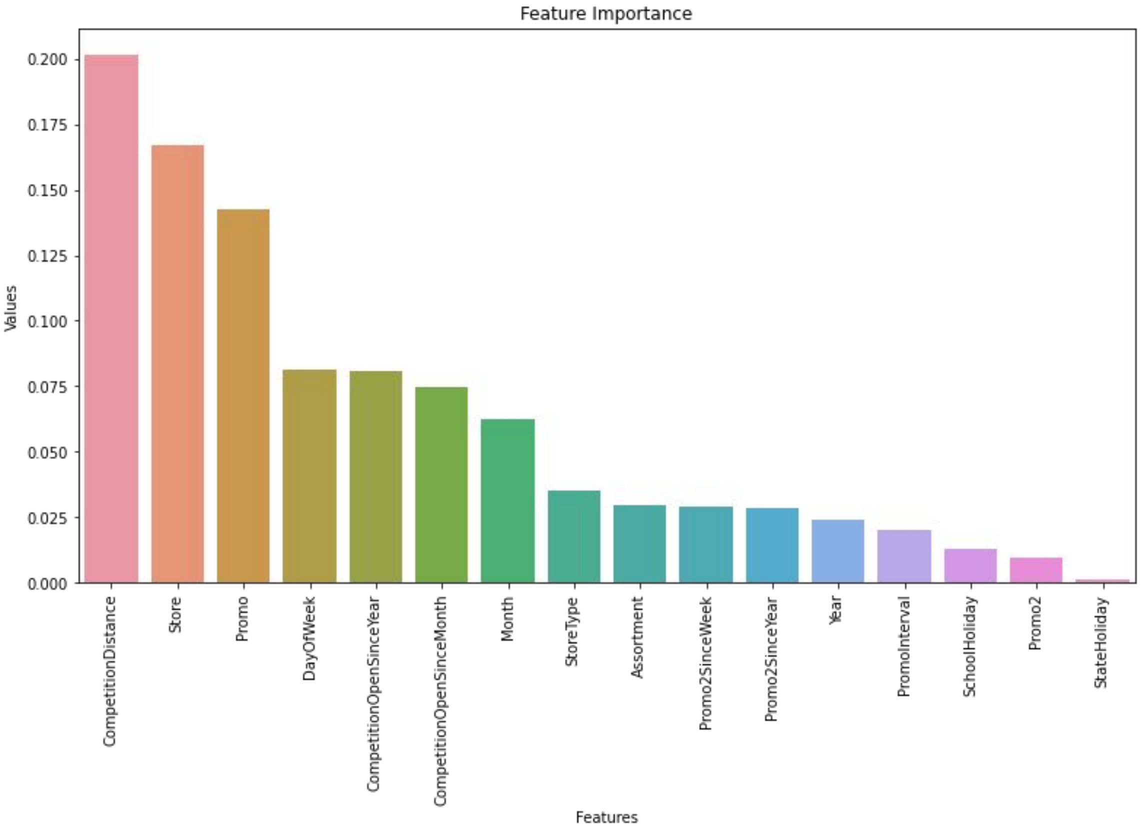


Observation - Random Forest regressor have the highest r2 score and lowest error as compared to other models

Important Features

Top 5 important Features are:

1. CompetitionDistance
2. Store
3. Promo
4. DayOfWeek
5. CompetitionOpenSinceYear



Conclusion

As competition is very high, promo codes can help them to boost their sales. We have observed that sales also depend upon the store type. Store model B has the highest number of sales out of all store models like A, B, C, and D. Sales was also depending upon the days of the week, as Monday and Friday received the highest number of sales. Also, the assortment has a role in sales and assortment level b has the highest sales.

We have seen a huge difference in sales when the Promo code was available in the store.

Also, most of the stores are closed on Holidays, so state holidays do not affect sales

Thank you