Capstone Project

On

Retail Sales Prediction

by

1. Subodh Shankar Dooganavar
2. Kasmin Talukdar

Problem Statement

- Rossmann operates over 3,000 drug stores in 7 European countries. Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance.

- With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

- You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

Retail Sales Prediction

- Demand for product or service is not constant, it changes with respect to time. But to maintain the demand and supply balance it is very important to understand the demand of product or service in the future.

- Sales prediction is a process of estimating demand of sales of a particular product or service over a period of time.

- Sales prediction not only helps in balancing the supply chain but also helps in making future business strategies like budgets, hiring, incentives, goals, acquisitions and various other growth plan.

Data Summary

Details of dataset – 1) **Rossmann Stores Data.csv** - historical data including Sales
2) **store.csv** - supplemental information about the stores

Description of data fields

- Id - an Id that represents a (Store, Date) duple within the test set
- Store - a unique Id for each store
- Sales - the turnover for any given day (this is what you are predicting)
- Customers - the number of customers on a given day
- Open - an indicator for whether the store was open: 0 = closed, 1 = open
- 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

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

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

•CompetitionDistance - distance in meters to the nearest competitor store

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

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

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

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

•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

Rossmann Stores Data.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1017209 entries, 0 to 1017208
Data columns (total 9 columns):
 #   Column         Non-Null Count    Dtype 
---  ------         --------------    ----- 
 0   Store          1017209 non-null  int64 
 1   DayOfWeek      1017209 non-null  int64 
 2   Date           1017209 non-null  object
 3   Sales          1017209 non-null  int64 
 4   Customers      1017209 non-null  int64 
 5   Open           1017209 non-null  int64 
 6   Promo          1017209 non-null  int64 
 7   StateHoliday   1017209 non-null  object
 8   SchoolHoliday  1017209 non-null  int64 
dtypes: int64(7), object(2)
memory usage: 69.8+ MB
```

store.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1115 entries, 0 to 1114
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Store                      1115 non-null   int64  
 1   StoreType                  1115 non-null   object 
 2   Assortment                 1115 non-null   object 
 3   CompetitionDistance        1112 non-null   float64
 4   CompetitionOpenSinceMonth  761 non-null    float64
 5   CompetitionOpenSinceYear   761 non-null    float64
 6   Promo2                     1115 non-null   int64  
 7   Promo2SinceWeek            571 non-null    float64
 8   Promo2SinceYear            571 non-null    float64
 9   PromoInterval              571 non-null    object 
dtypes: float64(5), int64(2), object(3)
memory usage: 87.2+ KB
```

Merged Dataset

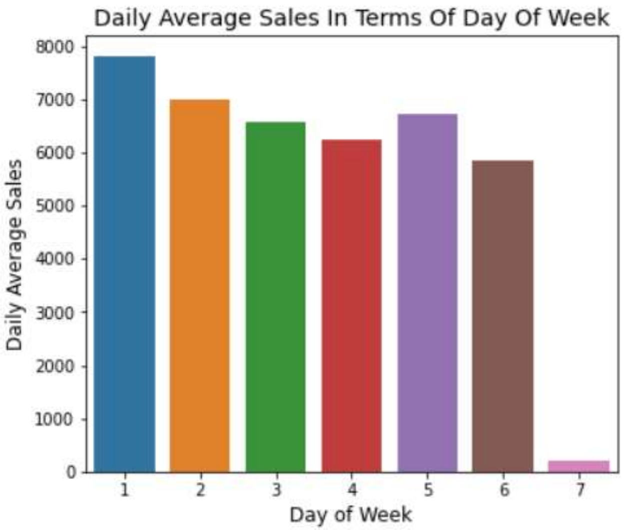
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 844392 entries, 0 to 1017190
Data columns (total 23 columns):
 #   Column                     Non-Null Count   Dtype
---  ------                     --------------   -----
 0   Store                      844392 non-null  int64
 1   DayOfWeek                  844392 non-null  int64
 2   Date                       844392 non-null  object
 3   Sales                      844392 non-null  int64
 4   Customers                  844392 non-null  int64
 5   Open                       844392 non-null  int64
 6   Promo                      844392 non-null  int64
 7   StateHoliday               844392 non-null  object
 8   SchoolHoliday              844392 non-null  int64
 9   StoreType                  844392 non-null  object
 10  Assortment                 844392 non-null  object
 11  CompetitionDistance        842206 non-null  float64
 12  CompetitionOpenSinceMonth  575773 non-null  float64
 13  CompetitionOpenSinceYear   575773 non-null  float64
 14  Promo2                     844392 non-null  int64
 15  Promo2SinceWeek            421085 non-null  float64
 16  Promo2SinceYear            421085 non-null  float64
 17  PromoInterval              421085 non-null  object
 18  Date-time                  844392 non-null  datetime64[ns]
 19  year                       844392 non-null  int64
 20  month                      844392 non-null  int64
 21  day                        844392 non-null  int64
 22  current_week_number        844392 non-null  int64
dtypes: datetime64[ns](1), float64(5), int64(12), object(5)
```

Null Values of merged dataset

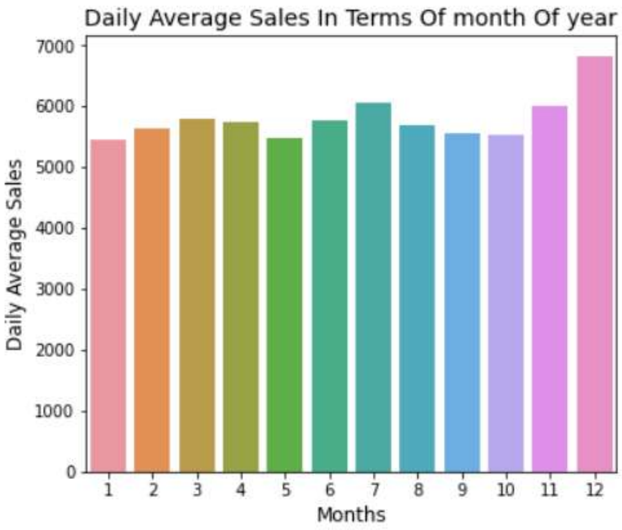
```
Store                     0
DayOfWeek                 0
Customers                 0
Promo                     0
StateHoliday              0
SchoolHoliday             0
year                      0
month                     0
day                       0
current_week_number       0
StoreType                 0
Assortment                0
CompetitionDistance    2186
Promo2                    0
competition_open          0
promo_2_open              0
IsPromo2Month             0
dtype: int64
```

Exploratory Data Analysis

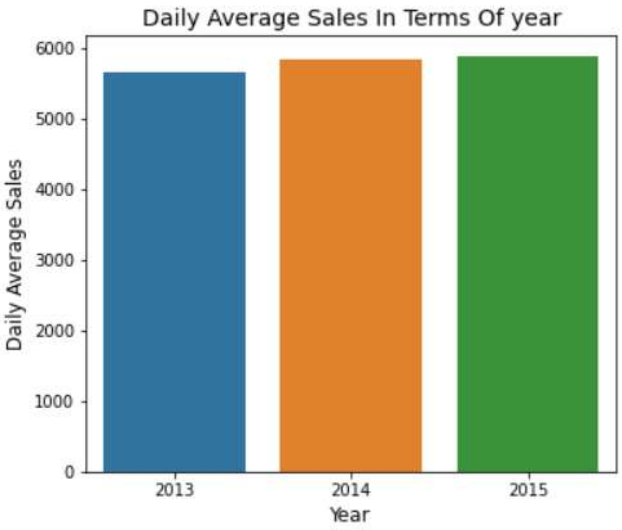
Daily average sales v/s Day of week



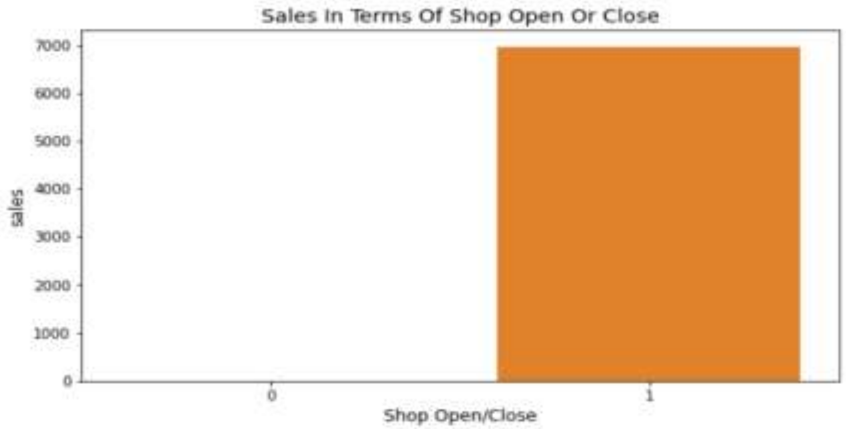
Daily average sales v/s Months



Daily average sales v/s Year

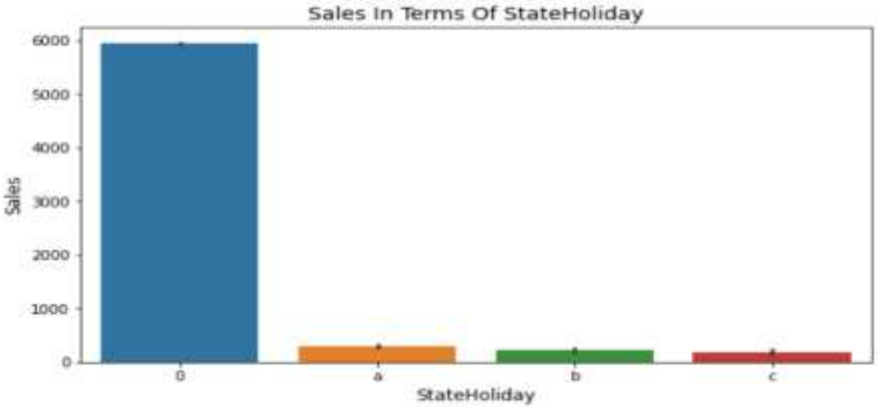


Sales v/s Shop open or close

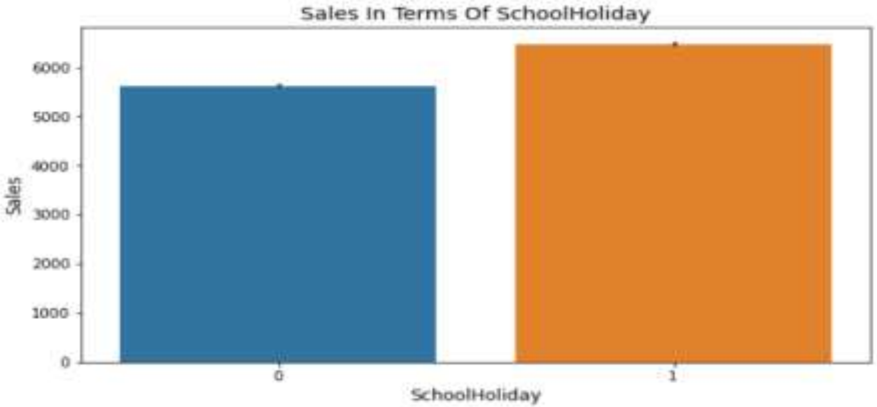


Sales v/s Promo availability

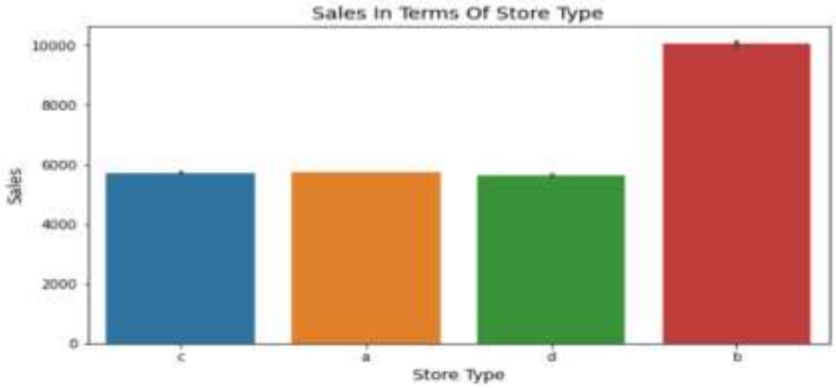
Sales v/s State holiday



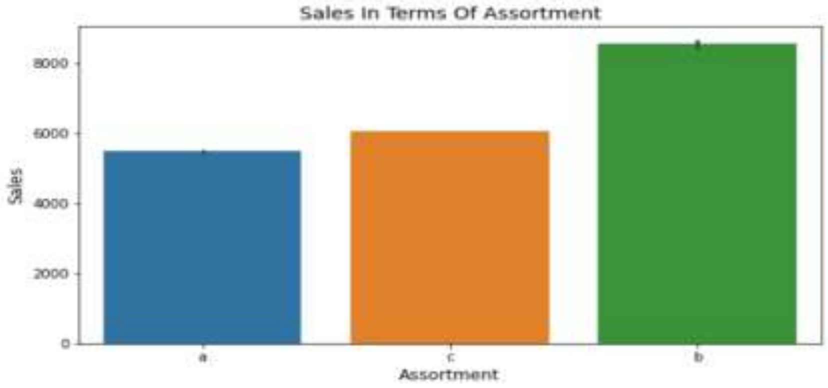
Sales v/s School holiday



Sales v/s Store type

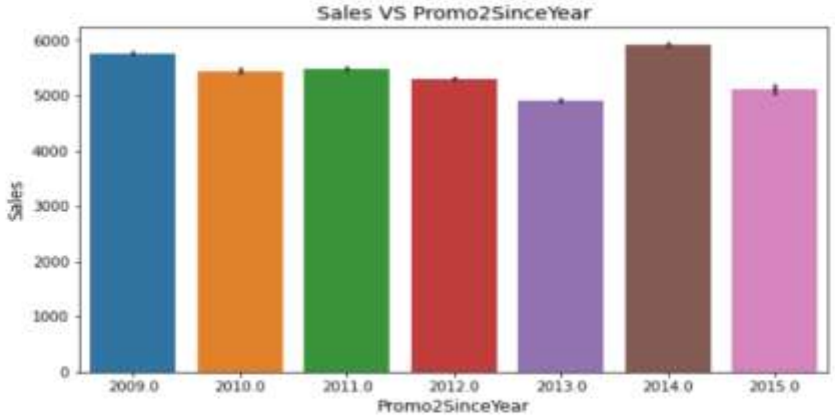


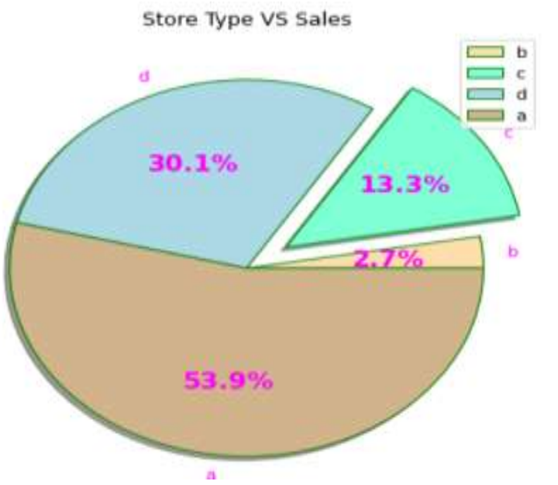
Sales v/s Assortment

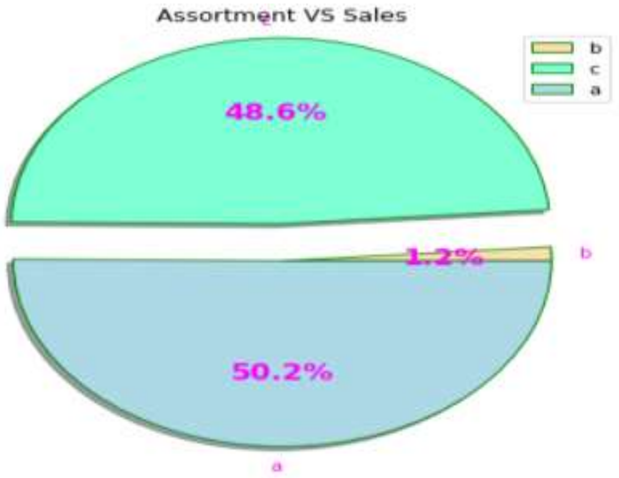


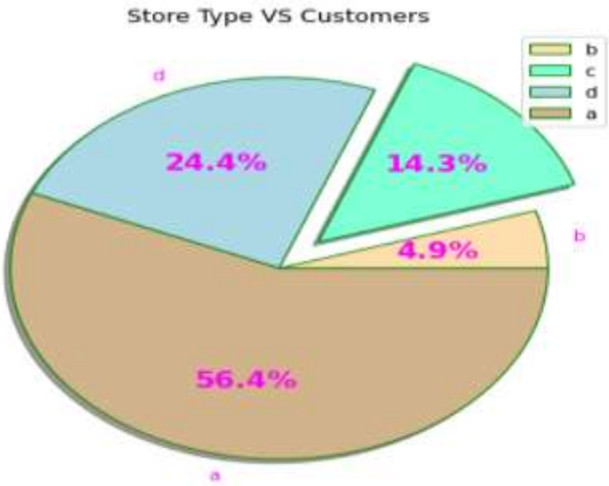
Sales v/s Promo2

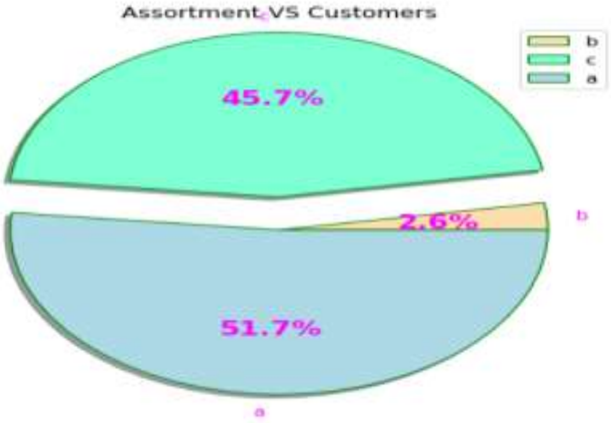
Sales v/s Promo2 since year



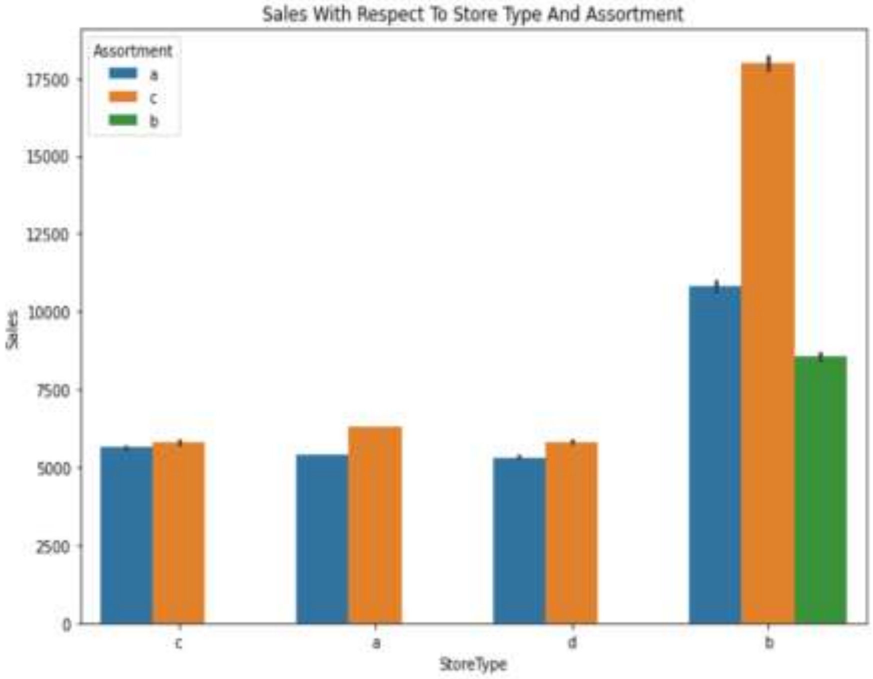

Store Type VS Sales
b
c
d
a
30.1%
13.3%
2.7%
53.9%


Assortment VS Sales
b
c
a
48.6%
1.2%
50.2%

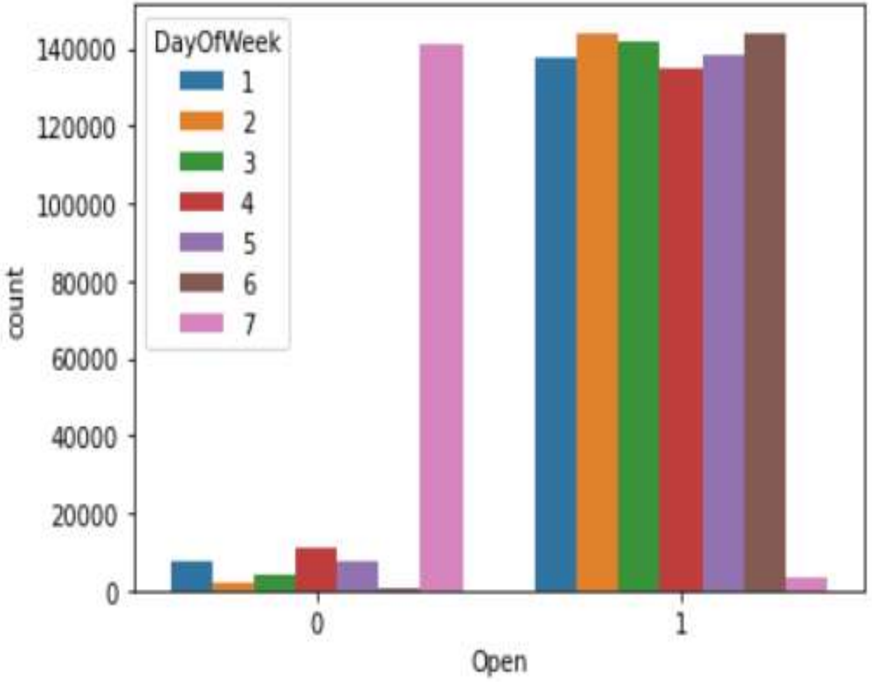

Store Type VS Customers
b
c
d
a
24.4%
14.3%
4.9%
56.4%


Assortment VS Customers
b
c
a
45.7%
2.6%
51.7%

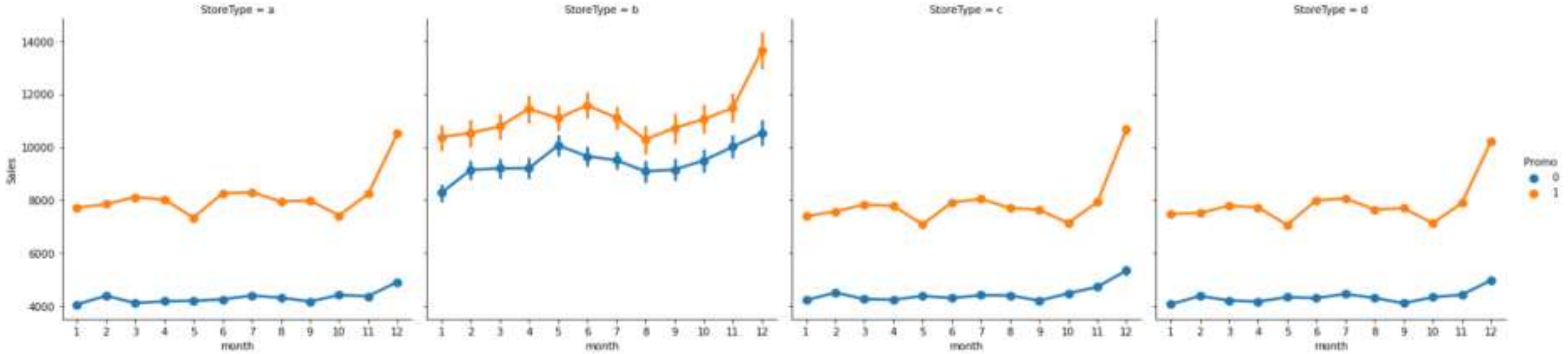
Sales with respect to store type and assortment



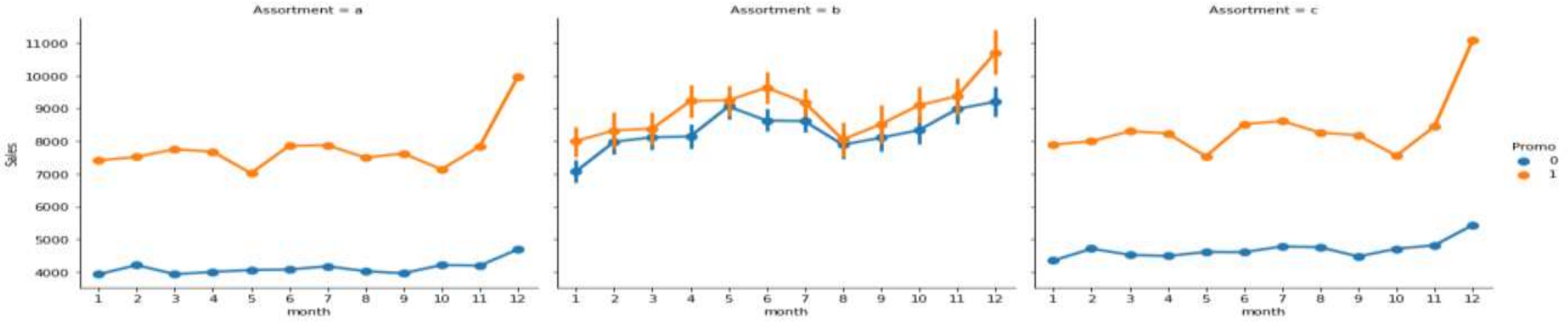
Count of shops open and closed



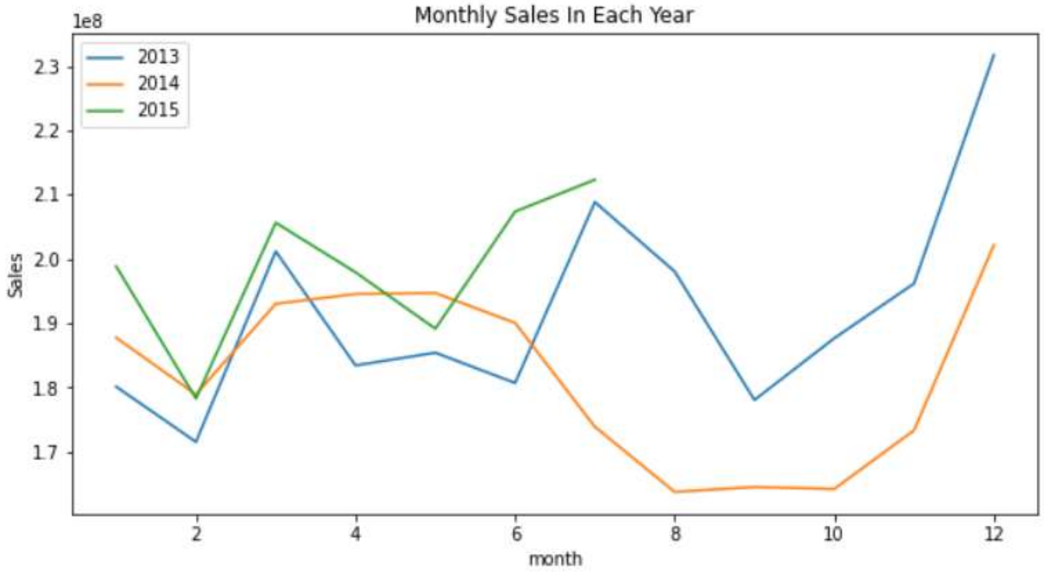
Monthly sales in terms of Store type and Promo



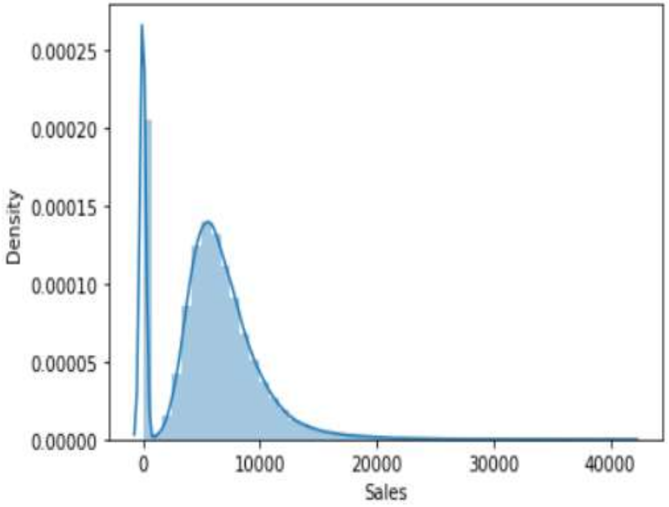
Monthly sales in terms of Assortment and Promo



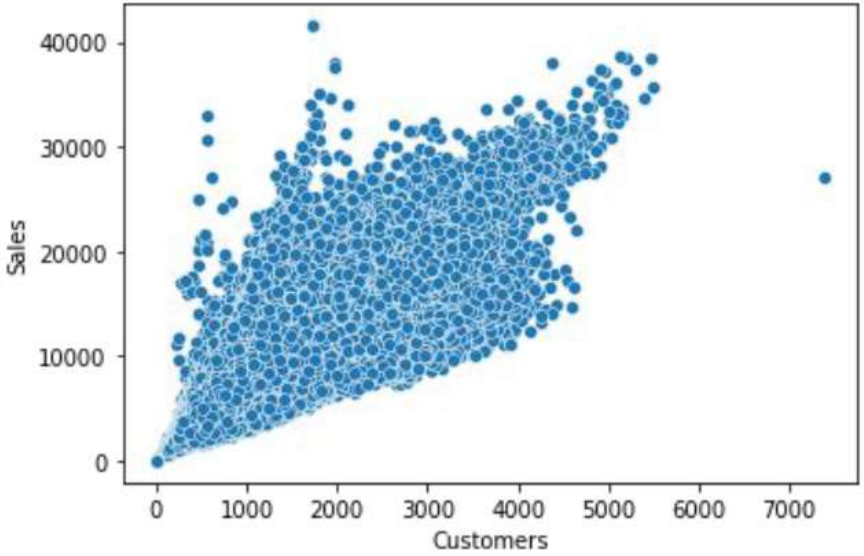
Monthly sales in each year



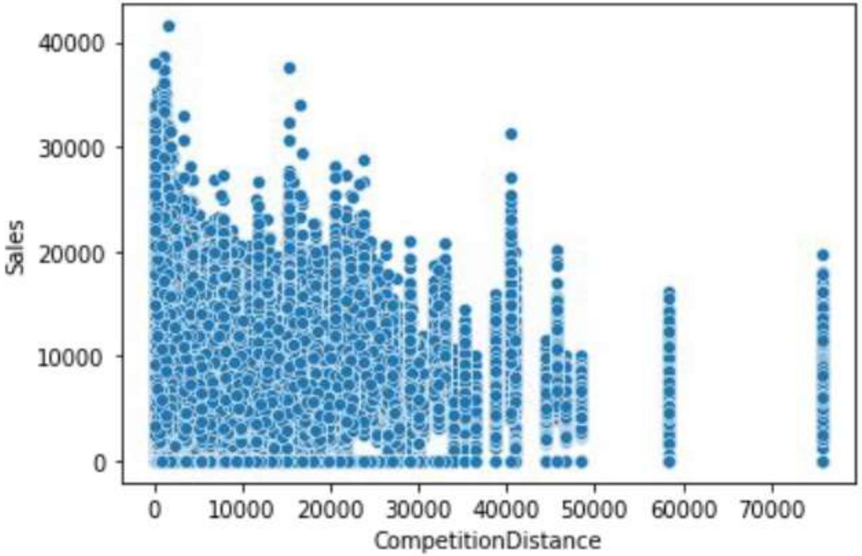
Sales Density



Pair plot of Sales v/s Customers



Pair plot of Sales v/s Competition distance



Insights from EDA

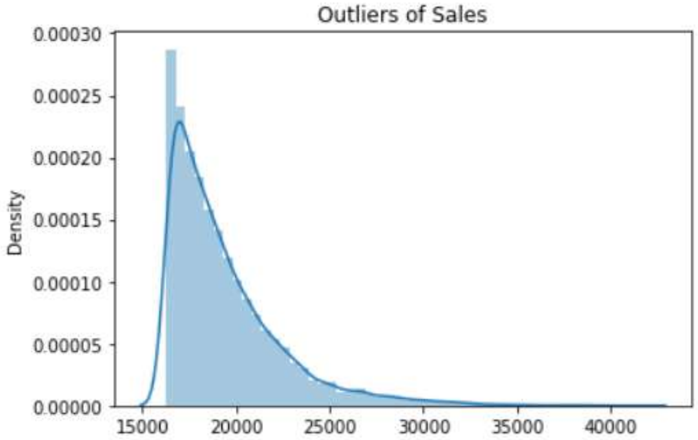
- Sales are high on Monday, December has the highest sales.
- 0 represents shop is basically closed so there is no sale on that day. Sales are pretty high when promo is available.
- More stores were open on School Holidays than on State Holidays and hence had more sales than State Holidays.
- On an average store type B and assortment type b had the highest sales.
- With Promo2, slightly more sales were seen without it which indicates there are many stores not participating in promo.
- Earlier it was seen that the store type 'b' had the highest sales on an average because the default estimation function to the bar plot is mean. But upon further exploration it can be clearly observed that the highest sales belonged to the store type 'a' due to the high number of type 'a' stores in our dataset.
- The drop in sales indicates the 0 sales accounting to the stores temporarily closed due to refurbishment.

Data Manipulation and Feature Engineering

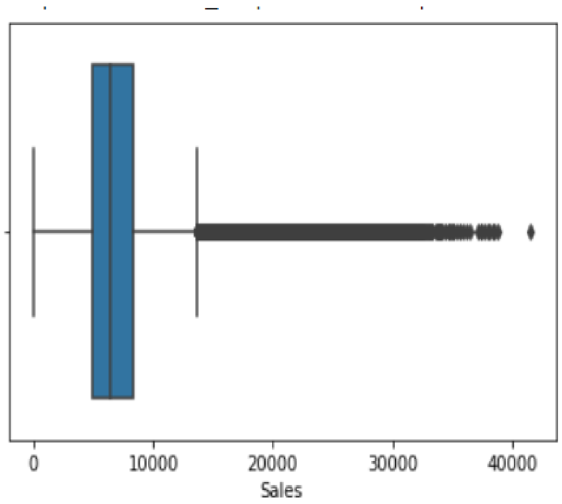
- Extracting current date, month, year, week number from date column.
- Since there is no sales when the shops are closed, we removed all the observations when the store is closed.
- Combine CompetitionOpenSinceMonth, CompetitionOpenSinceYear to give "competition _open“ which tells since how many months competition is open.
- Combine Promo2SinceWeek, Promo2SinceYear to give "promo_2_open“ which tells since how many months the shop is participating in promo2.
- Getting “IsPromo2Month” from promo_interval_open which tells is Promo2 open for a particular month or not.
- CompetitionDistance has some null values we will deal with it by filling null values with median of CompetitionDistance.

Outlier Detection

Using zscore



Using IQR (interquantile region)



Measure taken – Transformed the targeted variable to log scale.

Feature Scaling and One Hot Encoding

For numerical features – Scale the variables using Min-Max Scaling

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

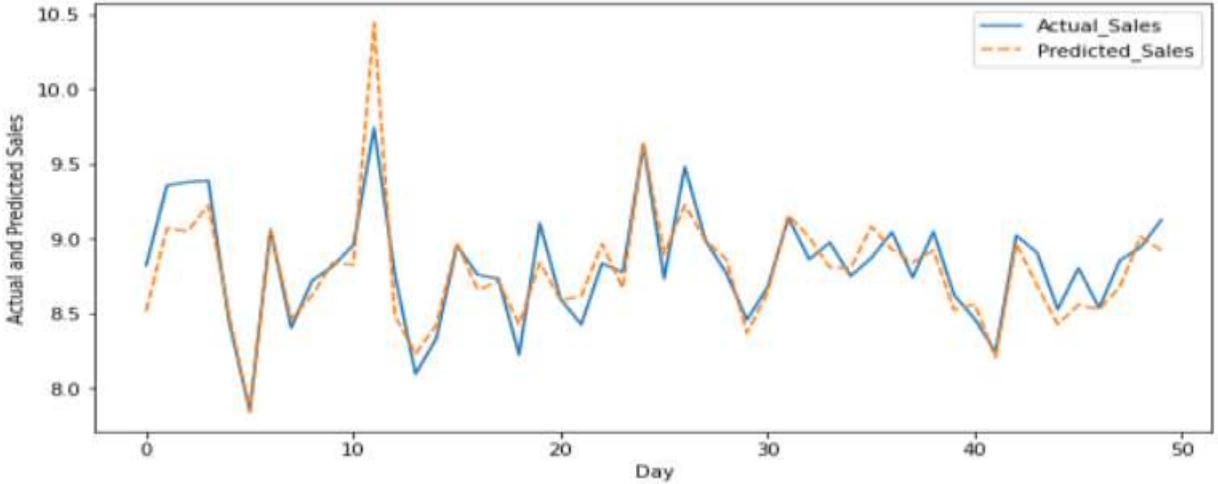
For categorial features – Apply one hot encoding on categorial variable to convert it to numerical.

Model Training

Different Models Used

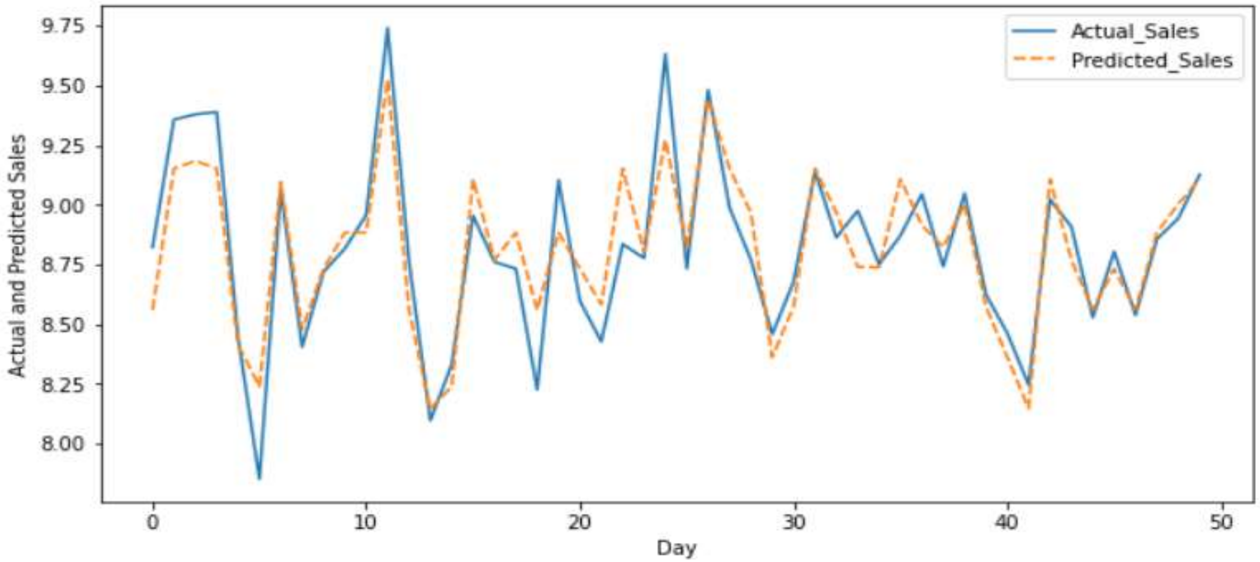
1. **Linear Regression**
2. **Decision Tree**
3. **Random Forest**
4. **Gradient Boosting**

Linear Regression



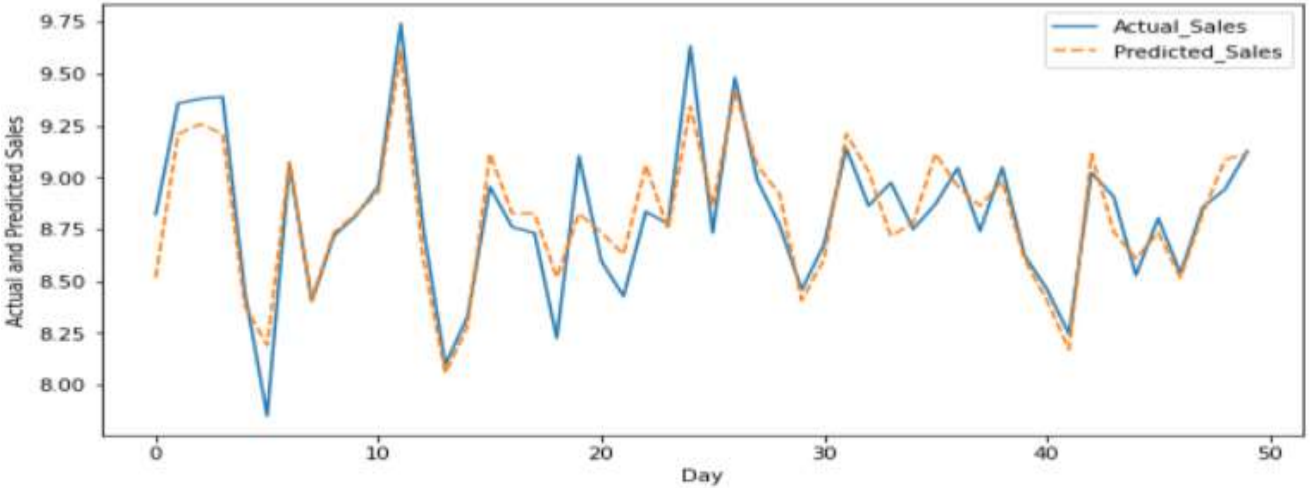
	Train Metrics	Train results	Test Metrics	Test results
0	Train_MAE	0.159226	Test_MAE	0.159264
1	Train_MSE	0.045045	Test_MSE	0.045348
2	Train_RMSE	0.212239	Test_RMSE	0.212950
3	Train_R2	0.750731	Test_R2	0.750107

Decision Tree



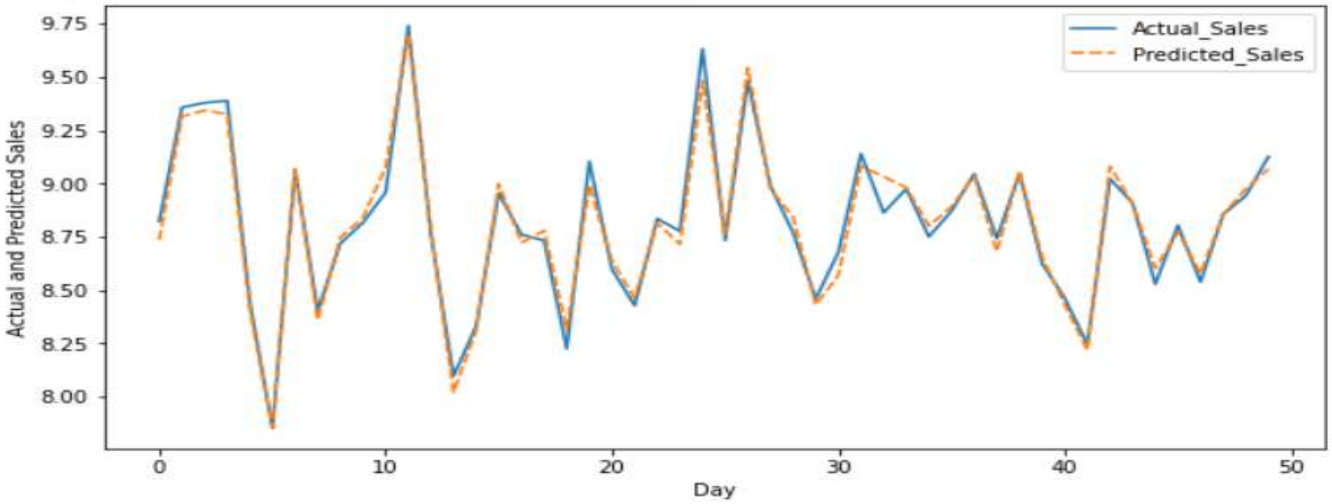
	Train Metrics	Train results	Test Metrics	Test results
0	Train_MAE	0.135781	Test_MAE	0.136305
1	Train_MSE	0.029625	Test_MSE	0.029912
2	Train_RMSE	0.172119	Test_RMSE	0.172952
3	Train_R2	0.836062	Test_R2	0.835166

Decision Tree (Hyperparameter Tuning)



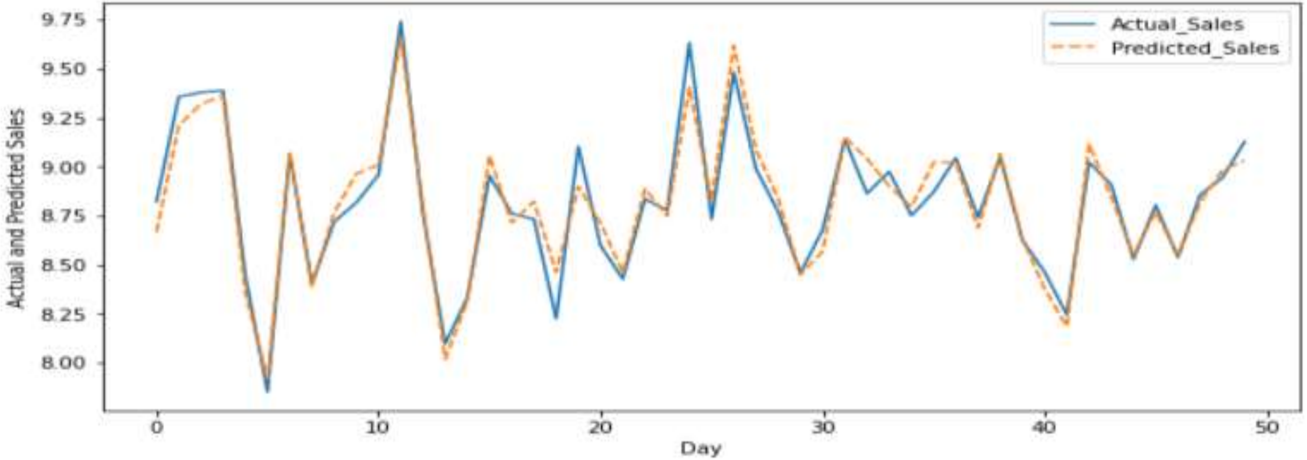
	Train Metrics	Train results	Test Metrics	Test results
0	Train_MAE	0.129045	Test_MAE	0.129447
1	Train_MSE	0.026586	Test_MSE	0.026789
2	Train_RMSE	0.163053	Test_RMSE	0.163674
3	Train_R2	0.852878	Test_R2	0.852375

Random Forest



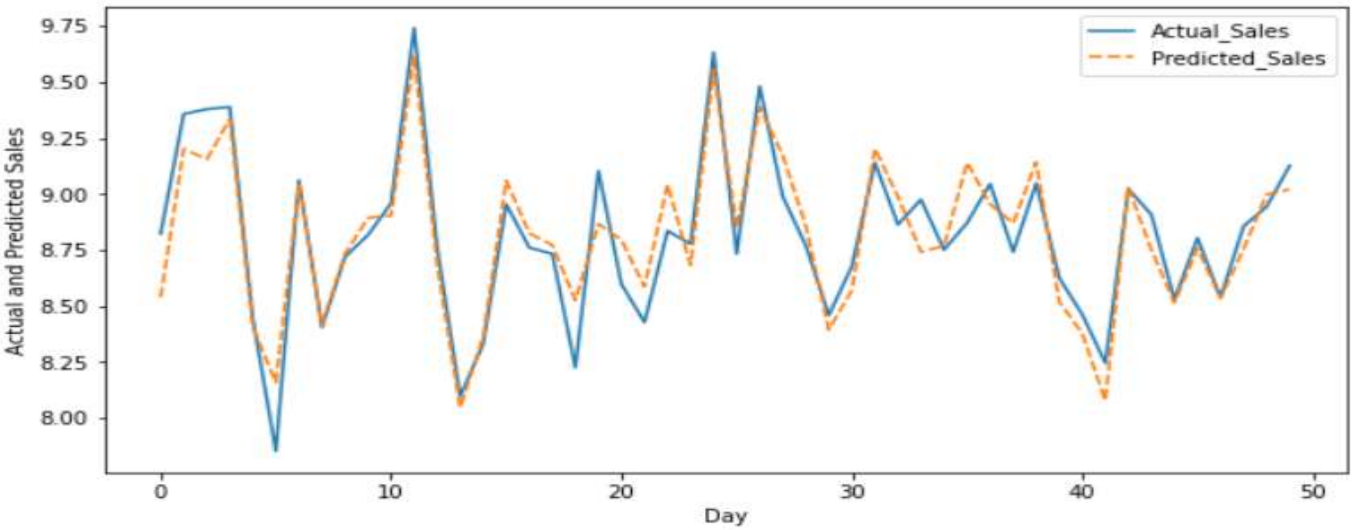
	Train Metrics	Train results	Test Metrics	Test results
0	Train_MAE	0.019115	Test_MAE	0.049246
1	Train_MSE	0.000709	Test_MSE	0.004439
2	Train_RMSE	0.026627	Test_RMSE	0.066626
3	Train_R2	0.996076	Test_R2	0.975538

Random Forest (Hyperparameter Tuning)



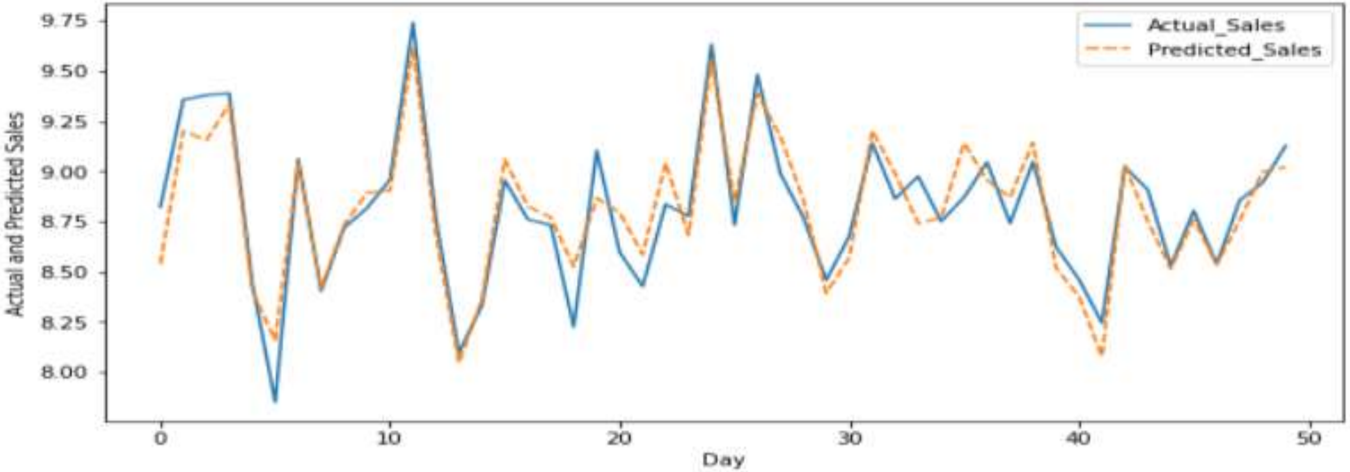
	Train Metrics	Train results	Test Metrics	Test results
0	Train_MAE	0.066516	Test_MAE	0.069251
1	Train_MSE	0.007676	Test_MSE	0.008455
2	Train_RMSE	0.087615	Test_RMSE	0.091952
3	Train_R2	0.957521	Test_R2	0.953407

Gradient Boosting



	Train Metrics	Train results	Test Metrics	Test results
0	Train_MAE	0.110526	Test_MAE	0.110699
1	Train_MSE	0.019615	Test_MSE	0.019683
2	Train_RMSE	0.140054	Test_RMSE	0.140297
3	Train_R2	0.891455	Test_R2	0.891534

Gradient Boosting (Hyperparameter Tuning)

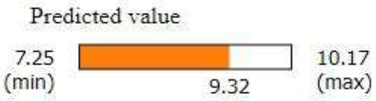


	Train Metrics	Train results	Test Metrics	Test results
0	Train_MAE	0.110526	Test_MAE	0.110699
1	Train_MSE	0.019615	Test_MSE	0.019683
2	Train_RMSE	0.140054	Test_RMSE	0.140297
3	Train_R2	0.891455	Test_R2	0.891534

Random Forest (Feature Importance)

Using LIME

```
Intercept 8.660910259249262
Prediction_local [9.50320815]
Right: 9.31539642454736
```



Feature	Value
Customers	0.20
StoreType4	0.00
Promo	1.00
CompetitionDistance	0.20
promo_2_open	0.00

After observing many observations it is observed that the following features are important Customer, Promo, Assortment2 and StoreType4.

Using ELI5

Top 5 features from ELI5

Weight	Feature
0.7533 ± 0.0041	x1
0.0571 ± 0.0011	x7
0.0485 ± 0.0035	x27
0.0381 ± 0.0013	x0
0.0366 ± 0.0011	x2

```
('Customers',
 'CompetitionDistance',
 'StoreType4',
 'Store',
 'Promo',
```

Conclusion

Sales Prediction helps in making future business strategies like budgets, hiring, incentives, goals, acquisitions and various other growth plan. In this project we analyzed more than one thousand stores for sales prediction. After analysing we conclude some important observations as follows

1. Stores which are running promo have more sales.

2. The State Holiday affects adversely to sales while school holiday affects positively to sales.

3. Store type B though being few in number had the highest sales average. The reasons include all three kinds of assortments specially assortment level b which is only available at type b stores and being open on Sundays as well.

4. With increase in competition distance sales decrease. This may be because the store with low competition distance indicates that the store is in busy place.

Challenges faced

- First of all the dataset involves time series. Again all the factors which we considered may not be effective for a long period of time. So our prediction may not give same accuracy as time changes.

- The major challenge would be the computational time and RAM needed to work upon such a dataset in a cloud environment.