Analyzing how competition and promotions influence sales

Author: Apurva Gupta

Given: Store data and sales data files

Step 1: Data Preprocessing

Sales Data

a. I removed the data where sales were 0.0 and store was closed. 172817 rows were dropped. Such rows won't affect our model.
b. There are 45 records which have NA values in 'Open' column but sales greater than 0. This means that store was open on that day. Therefore, I imputed 1.0 in 'Open' column for those records.
c. There are 103 records with unknown 'SchoolHoliday' column. I imputed them with mode value of data.

Store Data

a. I imputed CompetitionDistance with mean. CompetitionOpenSinceMonth and CompetitionOpenSinceYear with mode.
b. Promo2SinceWeek, Promo2SinceYear and PromoInterval were null for those rows where Promo2 = 0. Therefore, I replaced null values with 0,0.0 and Not Applicable respectively.

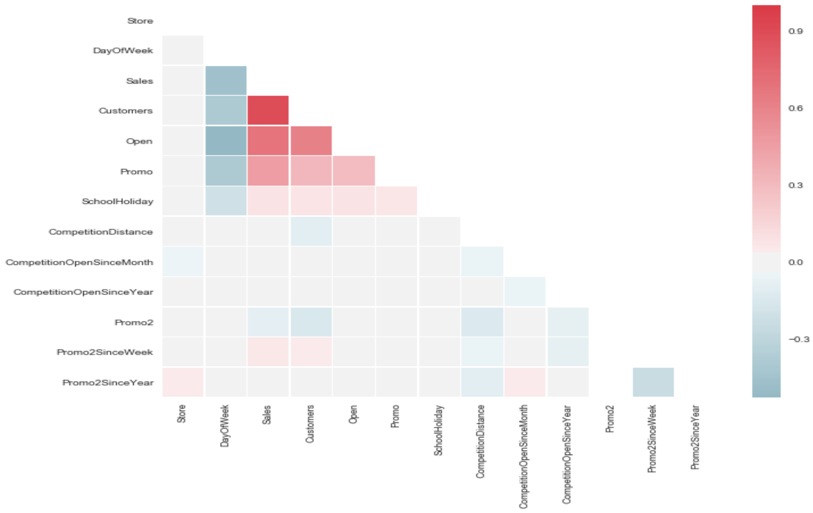
Now, the data doesn't contain any missing values.

Step2: Merge both the files on Store ID.

Step3: Create dummy variables for all the categorical columns: StateHoliday, PromoInterval, StoreType and Assortment.

Step4: Check correlations between features

We should remove the highly correlated features before model building.



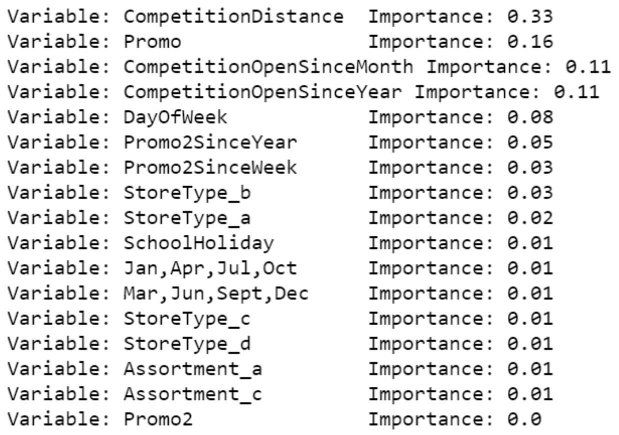
We can see that there is a strong correlation between Customers and sales(~0.9). We will remove 'customers' before building model.

Step 5: **Feature Selection:** Remove constant and quasi constant features

a. 'Open' is constant through all the rows. Therefore, I removed that column.
b. StoreID is unique identifier. Therefore, it cannot be used for prediction.
c. Date has a lot of unique values. Since we already have a column for holiday, we can ignore this column for prediction.

Step 6: Model Data using Random Forest Algorithm

a. Split data into train and test set (70-30 split)
b. Number of estimators used = 40
c. Check importance of features:



```
Variable: CompetitionDistance  Importance: 0.33
Variable: Promo                Importance: 0.16
Variable: CompetitionOpenSinceMonth Importance: 0.11
Variable: CompetitionOpenSinceYear Importance: 0.11
Variable: DayOfWeek            Importance: 0.08
Variable: Promo2SinceYear      Importance: 0.05
Variable: Promo2SinceWeek      Importance: 0.03
Variable: StoreType_b          Importance: 0.03
Variable: StoreType_a          Importance: 0.02
Variable: SchoolHoliday        Importance: 0.01
Variable: Jan,Apr,Jul,Oct      Importance: 0.01
Variable: Mar,Jun,Sept,Dec     Importance: 0.01
Variable: StoreType_c          Importance: 0.01
Variable: StoreType_d          Importance: 0.01
Variable: Assortment_a         Importance: 0.01
Variable: Assortment_c         Importance: 0.01
Variable: Promo2               Importance: 0.0
```

Our model finds that important variables used to predict the sales are: Competition Distance, Promo, Competition Open Month and Year, Day Of week.

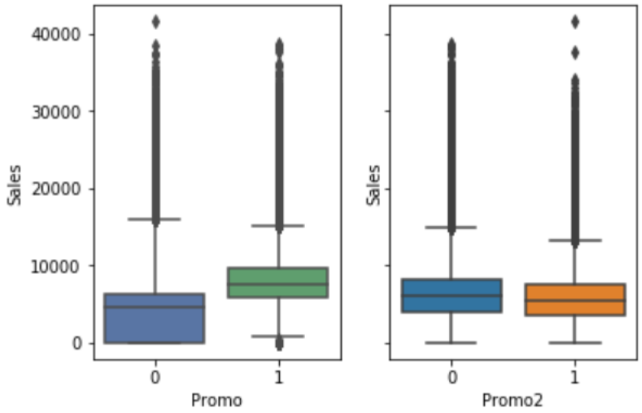
d. Performance Evaluation

R^2 : 0.838
Adjusted R^2 : 0.838
RMSE: 1246.49
MAE: 828.74

We can further improve the performance by different boosting algorithms like XGBoost.

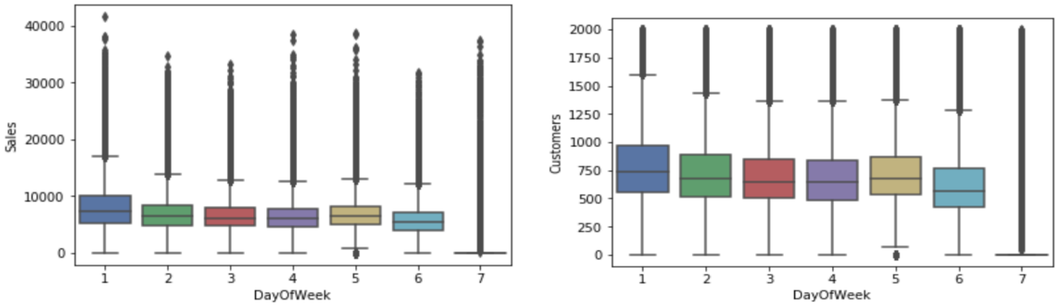
Exploratory Data Analysis

1. Effect of promotions on Sales



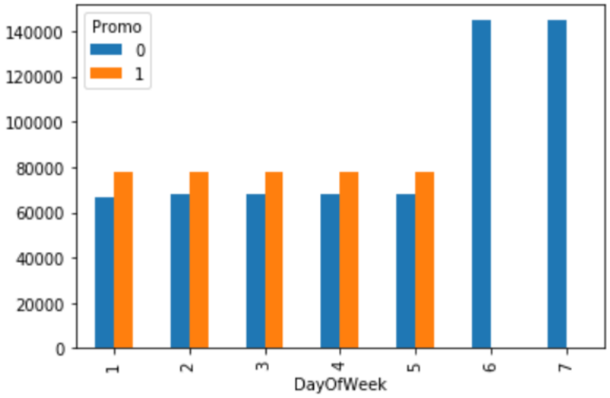
The box plots show that when there is an active Promo, Sales are more. Whereas, we can see that there is no effect of Promo 2 on our sales. The continuing and consecutive promotion do not cause any increase in sales. Our model also predicted Promo as one of the important variables and Promo2 as not an important parameter.

2. Impact of Day of Week on Sales and number of Customers



This gives us a very interesting insight. On Sundays, Walmart have very less Sales and customers. Let us try to find out the cause for this.

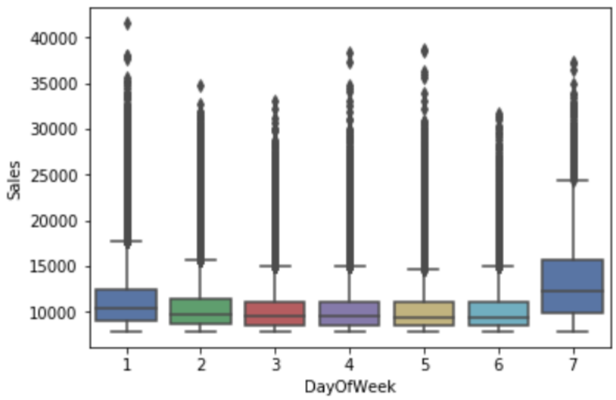
3. Day of Week vs Promo



On Saturday and Sundays, there are no promotions offered by Walmart. This could be the reason that there are less customers on Sunday.

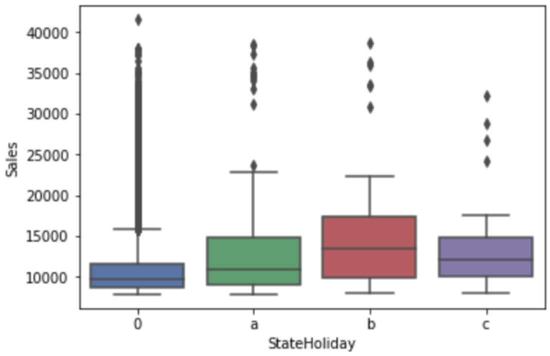
4. To find out that what can cause large sales, we will find the sales in 4th quantile and Analyze that.

   a. Day of Week vs Sales in 4th Quantile



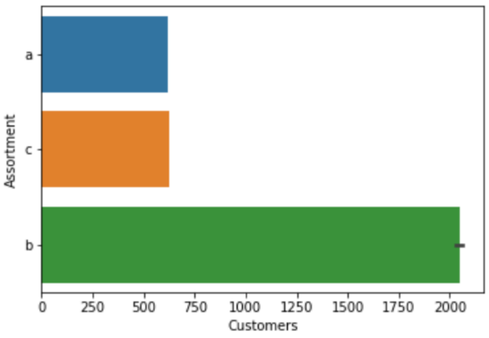
This gives an interesting insight. Looking at sales in 4 th quantile, Sundays have large number of high valued sales compared to other days.

b. State Holiday vs Sales in 4th Quantile



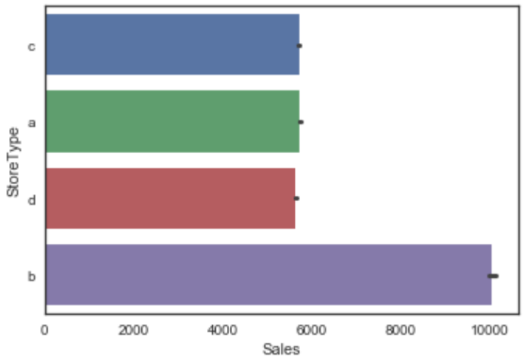
From the box plot, we can say that higher sales occur on Easter Holiday compared to other holidays.

5. Assortment vs Number of customers



We can observe that Customers who prefer assortment b are more than combined the customers who prefer a and c. This gives us an important insight that assortment extra level is very important for our customers.

6. StoreType vs Sales



Store b records the largest sales.

INSIGHTS TO ATTRACT CUSTOMERS AND INCREASE REVENUE

1. Promotions are an important parameter to attract customers. If there was promotion on that day, sales are higher as compared to no promotion days. Also, the visualizations show that customers and sales were less on weekends.
   Thus, giving out promotions on Weekends (especially Sunday) can attract more customers.
   Also, there could be a possibility that competing stores are running good promotions on weekends.
   We can analyze that data to find out that what kind of promotions can have higher customer footfall.
2. The highest valued sales happen on weekends. This shows that setting discounts on higher valued items on weekends could attract more people.
3. Promo 2 surprisingly doesn't show any impact on sales.
4. Number of customers visiting are more in store type B. For example, if we want to launch a new product, we can keep in stores of type B to attract more people.
5. Similarly, we can observe that assortment type B is highly preferred by customers.

Thus, our predicted sales model and various visualizations give us the different insights about the data. Adopting some of these can help us in increasing revenue and attracting customers.