**NextHikes**

Pharmaceutical Sales prediction across multiple stores

SUBMITTED BY

GEETIKA PANDEY

**Content**

**---------------------------------------------------------------------------------------------------**

1. **Problem Statement**
2. **Data Summary**
3. **Exploratory Data Analysis**
4. **Feature Engineering**
5. **Machine Learning Model Building**
6. **Challenges Faced**
7. **Conclusions**

## Problem Statement

-Rossmann operates over 3000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales up to six weeks in advance.

-Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.

-We are provided with historical sales data for 1,115 Rossmann stores.

-The task is to build a Machine Learning model to forecast the "Sales" column for the test set.

# DATA SUMMARY

## Details of Datasets Provided

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S. no.** | **Dataset** | **Variables** | **No. of Variables** | **No. of observations** |
| **1** | **Rossmann Stores Data**  **- gives historical data including Sales** | **store, day of week, date, sales,customers, open, promo,**  **state holiday, school holiday** | **9** | **1017209** |
| **2** | **store**  **- supplement information about the stores** | **store, storetype, assortment, competition distance, competition open since month, promo2, promo2since**  **week, promo2since year, promo interval** | **10** | **1115** |

**Important Variables**

* 1. **Sales - The turnover for any given day (this is what we are predicting).**
  2. **Open - An indicator for whether the store was open or**

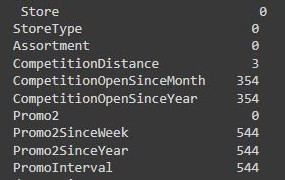
closed. 0 = closed, 1 = open.

* 1. **Store type - Differentiates between 4 different store models (a, b, c & d).**
  2. **Assortment - Describes an assortment level: a = basic, b = extra, c = extended.**
  3. **Promo - Indicates whether a store is running a promo on that day**
  4. **Promo2 - Promo2 is a continuing and consecutive promotion for some stores.**

## Important Variables(cont.)

* 1. **Store - A unique ID for each store.**
  2. **Customer - The number of customers on a given day.**
  3. **Competition Distance - Distance in meters to the nearest competitor store.**
  4. **Promo Interval - Describes the consecutive intervals Promo2 is started, naming the months the promotion is started new.**
  5. **Promo2Since [Year/week] - Describes the year and calendar week when the store started participating in Promo2.**

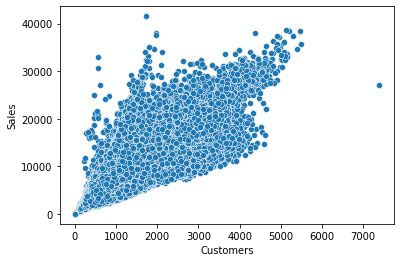
## Missing Values

The number of Missing values present in both the Datasets provided are :

**Rossmann Stores Data.csv store.csv**

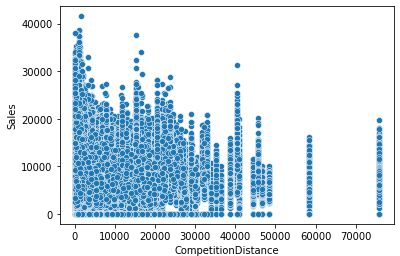
# EXPLORATORY DATA ANALYSIS

## Scatterplot - Sales and Customer

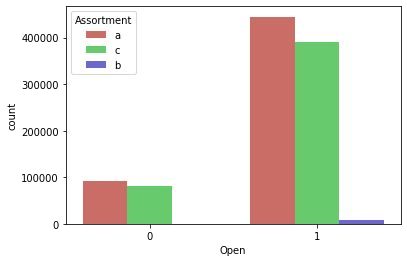
* **Scatterplot shows that the relation between Sales and customer is sort of linear. Sales are increasing with the number of Customers increasing which is pretty obvious.**

**Scatterplot - Sales and Competition Distance**

* **It can be observed that mostly the competitor stores weren't that far from each other and the stores densely located near each other saw more sales.**

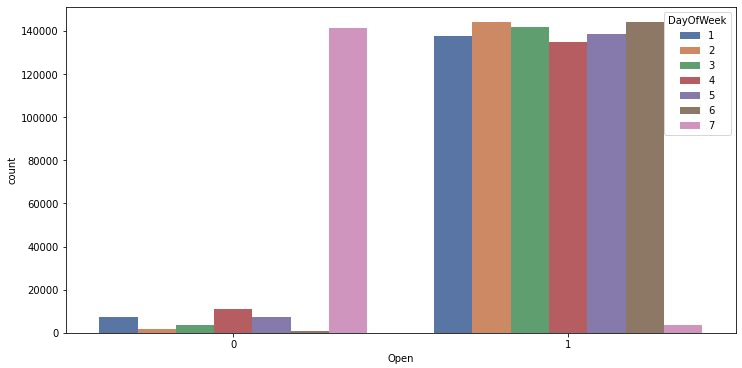


## Countplot - Store Open/Closed

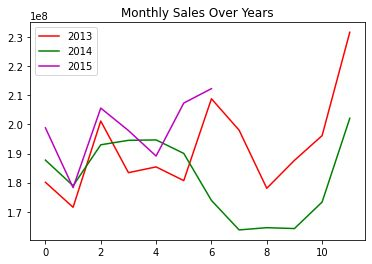


1. - closed
2. - opened

**Store Open/Closed over Day of Week**

* **This count plot clearly shows that majority of stores are closed on sunday. Some stores were also closed on other days of the week may be due to public holidays & refurbishment, as stores are usually closed on public holidays and are open during school vacations.**

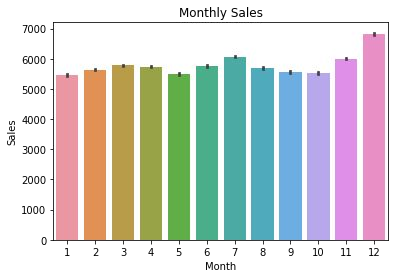
## Average Sales Over Year/Month

**- Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.**

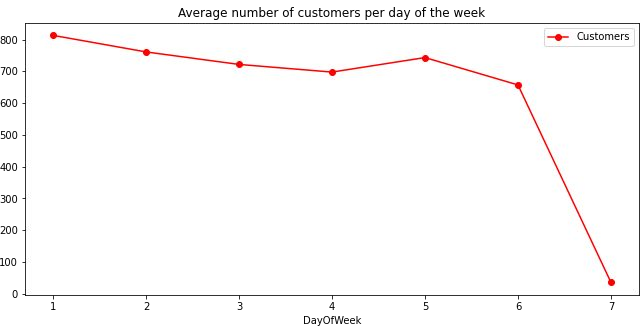
## Total Sales in successive Years



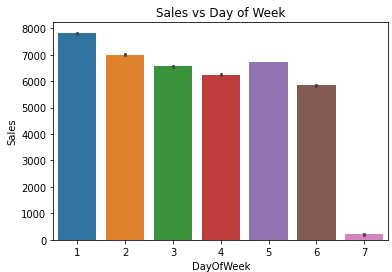
**Total Sales in Months**



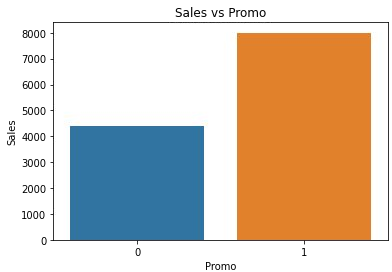
**Trend - Average Customer per Day of Week**



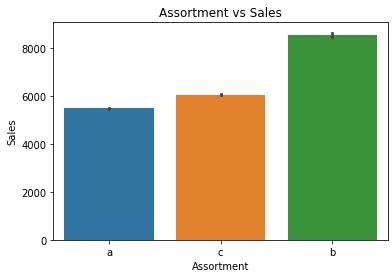
## Total Sales on Weekdays



**Sales VS Promo**



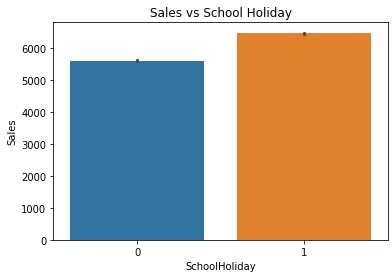
## Sales vs Assortment



a - basic b - extra

c - extended

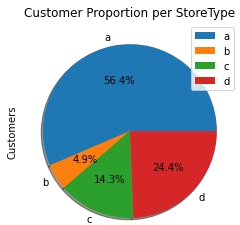
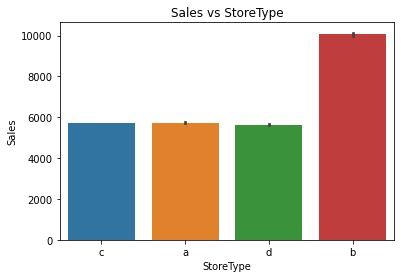
**Sales vs School Holiday**



1. - Not Holiday
2. - School Holidays

## Proportion of Store Types and Total

**Sales per Store Type**



# FEATURE ENGINEERING

## Replacing Missing Values

* **There are many Missing values in our Store Dataset.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Number of Missing values** | **Replaced with** | **Remark** |
| **Promo2SinceWeek** | **544** | **0** | **Not much Information given about the features.** |
| **Promo2SinceYear** | **544** |
| **PromoInterval** | **544** |
| **CompetitionOpenSinceMonth** | **354** | **Mode** |  |
| **CompetitionOpenSinceYear** | **354** |
| **CompetitionDistance** | **3** | **Median** | **As they are only 3.** |

## Changing Datatype of Columns

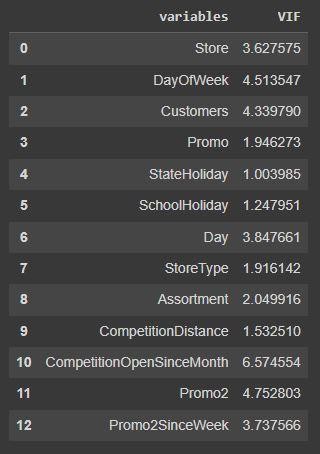
* + In Store Dataset – Store Type, Assortment and Promo Interval s of Object type.
  + Store Type and Assortment have values [a, b, c, d] and [a, b, c] respectively.
  + Changing these to Numerical values [0,1,2,3] and [0,1,2] respectively.

Promo Interval has values [0,'Feb, May, Aug, Nov', 'Jan, Apr, Jul, Oct', 'Mar, Jun, Sept, Dec']

* + Changing these to dummy variables**.**

## Correlation Heatmap

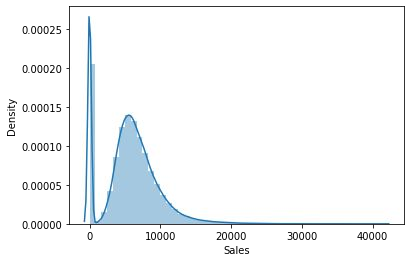
**Removing Multicollinearity**

* **Removing the features which is having VIF>10 because it will affect & interpret the result.**
* **VIF <= 10 is usually preferred as this can easily explain the variance of 90% i.e**

R-square becomes 90%.(VIF=1/1-R^2)

## Filtering Rows (Records)

* + **Filtering records where stores are closed as they won't generate any Sales.**
  + **Filtering records where stores has Sales equal to 0.**



# MACHINE LEARNING MODEL BUILDING

## Evaluation Metrics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ML Model predict sales for stores which are open and when there is some sales because there is no sales when store is closed.** | | | | | | |
|  | **Linear** | **Lasso** | **Ridge** | **Decision** | **Random** |  |
| **Regression** | **Regression** | **Regression** | **Tree** | **Forest** |  |
| **(OLS)** | **(L1)** | **(L2)** |  |  |  |
|  |  |  |  |  |  |
| **RMSE** | **1520.458** | **1521.955** | **1521.182** | **1433.262** | **1400.531** |  |
| **MAPE** | **15.919** | **15.938** | **15.926** | **15.477** | **15.221** |  |
| **R2** | **0.7607** | **0.7602** | **0.7605** | **0.7605** | **0.7969** |  |

## Model Selection

|  |
| --- |
| **By Looking at the evaluation metrics obtained on** |
| **implementing different sort of regression model, we** |
| **decided to go with the Random Forest model. The** |
| **maximum R^2 was seen in tuned Random Forest model** |

with the value 0.7969. It means our best accurate model is able to explain approx/almost 80% of variances in the datasets.

Based on our model Customer, store Type, Promo & Competition Distance are the most impactful features which are driving the sales more as compared to other features present in the dataset.

## Challenges Faced

* + **Understanding the meaning of some columns.**
  + **Handling Large amount of Sales Data.**
  + **Dealing with Null values, as there are many Null values.**
  + **Understanding the business model of Retail Sales that how they work.**
  + **Also, forming different graphs to show insights from the dataset and to summarize the information and communicate the results and trends to the reader successfully.**
  + **Dealing with Categorical columns to make them numerical for make use in ML model building.**
  + **Selecting appropriate Model to fulﬁll the purpose.**

## Conclusions

* **From the sales and customer scatterplot, the relationship is sort of linear Ie sales is increasing with number of customers increasing which is obvious.**
* **Stores with Assortment level ‘b’ has the highest sales.**
* **Approx. 50% stores are of type ‘a’. There are very few stores of type ‘b’.**
* **Store type ‘b’ has the highest sales and all other store types ‘a’, ’c’, ’d’ has nearly equal sales.**
* **December records the highest monthly sales. This may be due to Christmas and New Year.**
* **Sales is more when promos/offers are running on stores.**

### Recommendations from Analysis

#### There is seasonality involved. Hence, the stores should be encouraged to promote and take advantages of the holidays.

#### More stores should be encouraged for promos.

#### Store type ‘b’ should be increased in number

**Thank You**

**Hope model will help the Store managers to predict the daily sales accurately in advance.**

**Submitted by**

**GEETIKA PANDEY**

**(06-02-2024)**