Capstone Project

Rossman Retail Sales Prediction
(Machine Learning: Regression)

- **Team Members:**

- **Kumar Abhinav – kumarabhinavthakur274@gmail.com**
- **Saumya Dash – saumyadash9gmail.com**

Content

AI

- Introduction/Problem Statement
- Data Pipeline
- Exploring Dataset
- Attribute Information
- Data Cleaning and Handling
- Exploratory Data Analysis
- Machine Learning Algorithms for Modelling
- Model Performance and Evaluation Metrics
- Key Insights
- Conclusion/Recommendations

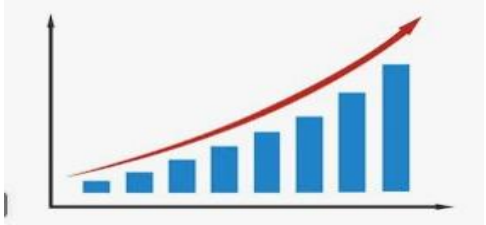
Introduction/ Problem Statement

Rossmann, one of the biggest drugstore chains in Europe, has more than 3000 outlets and employs about 56,200 people across seven different European countries. Rossmann store managers are currently required to forecast their daily sales up to six weeks in advance. Numerous factors, such as promotions, competition, school and state holidays, seasonality, and locality affect store sales.

The accuracy of the results can be highly variable because thousands of different managers are making sales predictions based on their own situations.

The task is to forecast the "Sales" column for the test set.

Why Sales Forecasting?



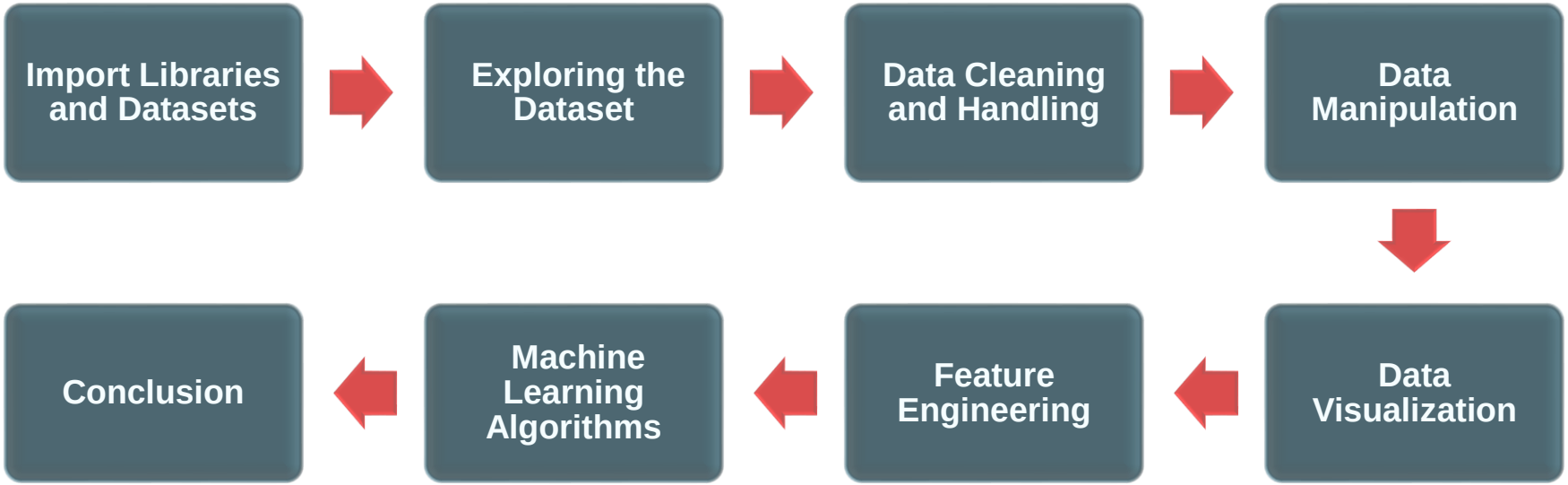
Definition

Sales forecasting is the technique of predicting demand or sales of a specific product over a predetermined time frame. Businesses use sales forecasts to estimate the revenue they will bring in over a specific period of time so they may create strong and effective business plans.

Objective

The revenue the firm expects to generate in the upcoming months has an impact on crucial decisions like budgets, hiring, incentives, objectives, acquisitions, and numerous other growth plans. Thus it's critical that these projections be accurate for the plans to be as successful as they are intended to be.

Data Pipeline



Exploring Dataset

Description	Rossman Stores Dataset	Stores Dataset
Information	Historical data including sales.	Supplemental information about the Stores.
Size	10,17,209 rows and 9 features	1,115 and 10 columns.
Column Names	Store, Day of Week, Date, Sales, Customers, Open, Promo, State Holiday, School Holiday	Store, Store Type, Assortment, Competition Distance, Competition Open Since Month, Competition Open Since Year, Promo2, Promo2 Since Week, Promo2 Since Year, Promo Interval.

Exploring Dataset

Description	Rossman Stores Dataset	Stores Dataset
Columns with Null Values	0 columns	6 columns: Competition Open Since Month, Competition Open Since Year, Promo2 Since Week, Promo2 Since Year, Promo Interval.
Datatype of the columns	Integer Data Type for all columns except for Date and State Holiday with Object Data Type	Integer and float datatype except for Store Type, Assortment and Object.

Attribute Information

Column Name	Description
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
State Holiday	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

Attribute Information

Column Name	Description
School Holiday	Indicates if the (Store, Date) was affected by the closure of public schools
Store Type	Differentiates between 4 different store models: a, b, c, d
Assortment	Describes an assortment level: a = basic, b = extra, c = extended
Competition Distance	Distance in meters to the nearest competitor store
Competition Open Since[Month/Year]	Gives the approximate year and month of the time the nearest competitor was opened

Attribute Information

Column Name	Description
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
Promo Interval	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

Data Cleaning and Handling

Treating the Null Values

Promo, Promo2 Since Week, Promo2 Since Year and Promo Interval has 544 null values which has been substituted with 0.

The Competition Distance has 3 missing data and is rightly skewed distribution thus is substituted with the Median of the data.

The Competition Open Since Month and Competition Open Since Year have 354 null values and are categorical columns, thus for substituting the null values we are using Mode of the data.

Data Cleaning and Handling

Data Manipulation

Open Column contains information about open or closed stores, indicated by the numbers 1 and 0, respectively. If the store is closed, there is probably no sales on that day. Data for closed stores are hence removed from the dataset.

There are 31460 instances where sales are zero. Since it makes up only 0.0067% of the entire dataset, the analysis won't be impacted even if it is dropped. Thus, records with no sales are removed from the dataset.

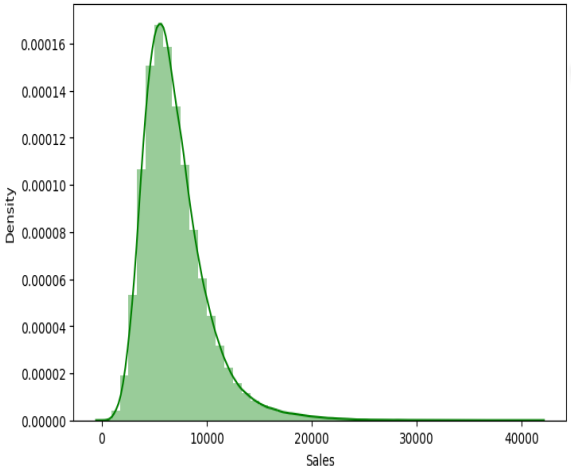
Month and Year is extracted from the Date Column to be used in our analysis. We have added the extracted Month and Year as a separate feature in our model.

We have added a column named average Store wise Sales for checking some trends in our analysis

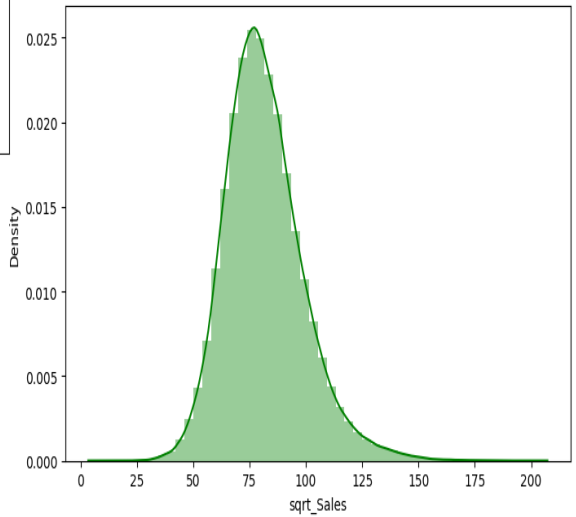
In the State Holiday Column we can see there are two zeroes in the Dataset. The first is a number, while the second is a string. Therefore, we have combined the two zeros into one, signifying there is no State Holiday on that specific day.

Data Visualization

Sales: Target Variable [Continuous Variable]



After Square Root Transformation

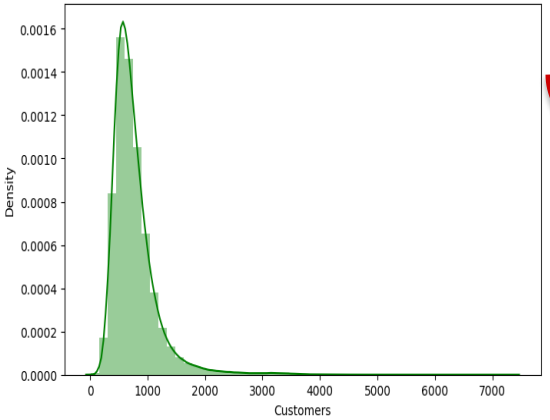


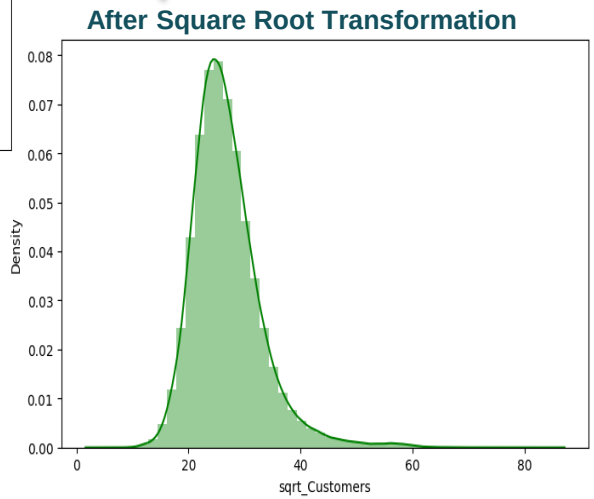
We can observe from the probability density curve for Sales
that the distribution is a little right-skewed.
We utilized a square root transformation function to adjust the symmetry and as can be
seen, the resulting graph is normally distributed.

Plot used: Distplot

Data Visualization

Customers [Continuous Variable]



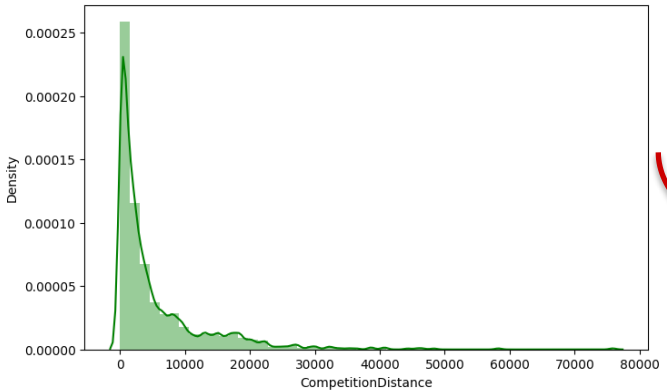


We can observe from the probability density curve for customers that the distribution is a little right-skewed. We utilized a square root transforma tion function to adjust the symmetry and as can be
seen, the resulting graph is normally distributed.

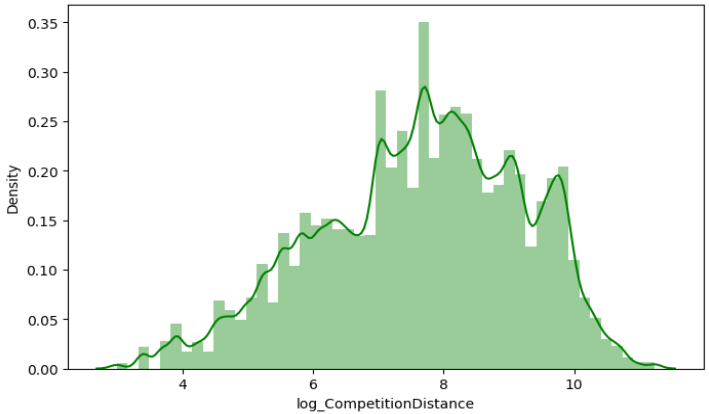
Plot used: Distplot

Data Visualization

Competition Distance [Continuous Variable]



After Log Transformation



We can observe from the probability density curve for Competition Distance that the distribution is Moderately right-skewed.
We utilized a Log transformation function to adjust the symmetry and as can be seen, the resulting graph is normally distributed.

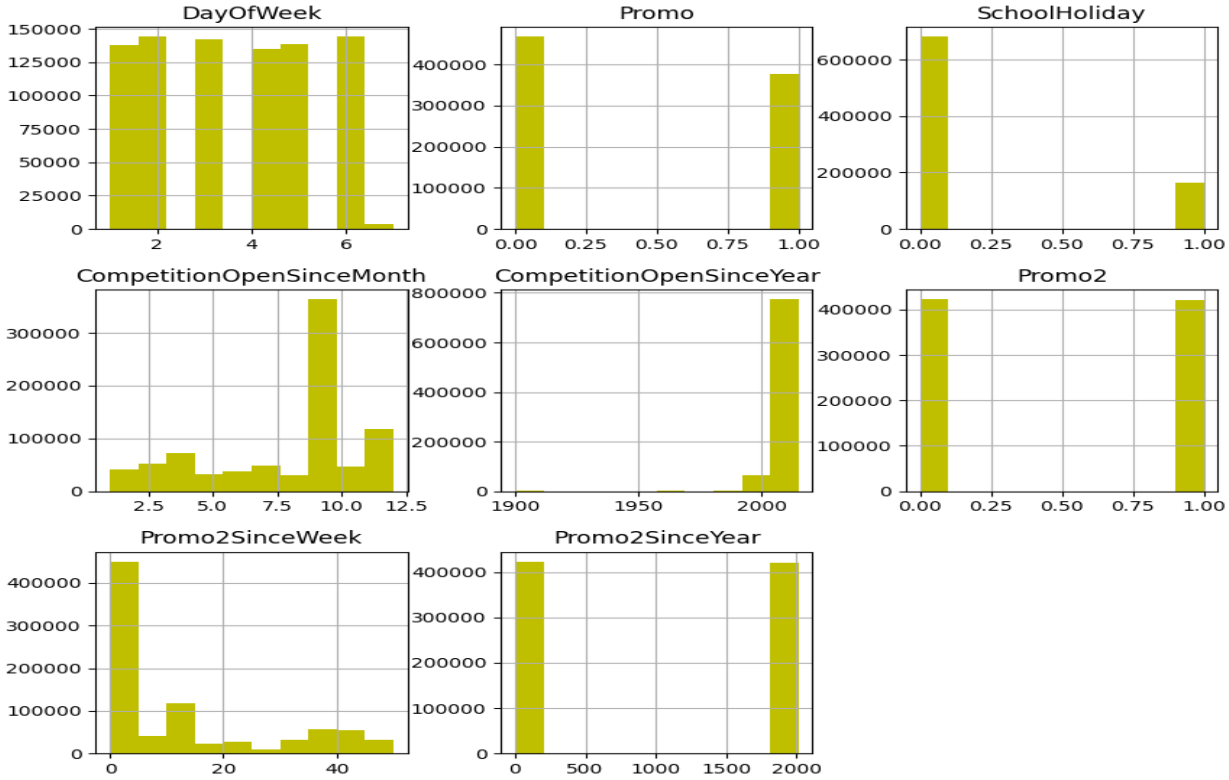
Plot used: Distplot

Data Visualization

Univariate Analysis:

Categorical Variables

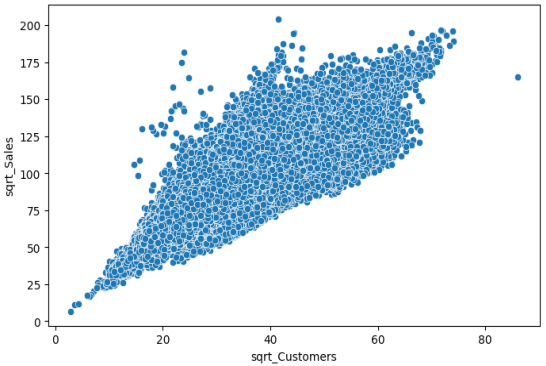
- Day Of Week
- Promo
- School Holiday
- Competition Open Since Month
- Competition Open Since Year
- Promo2
- Promo2 Since Week
- Promo2 Since Year



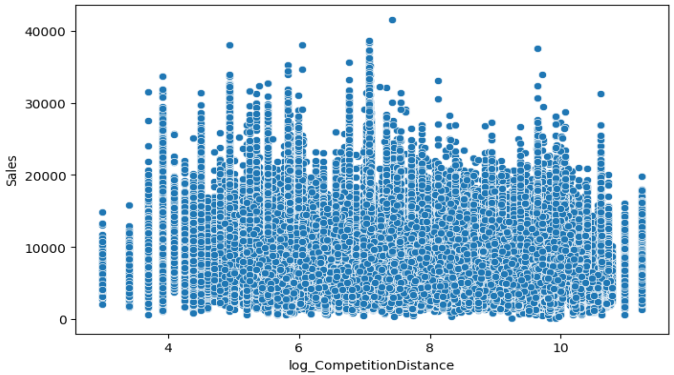
Plot used: Histogram

Data Visualization

Customers and Sales



Competition Distance and Sales

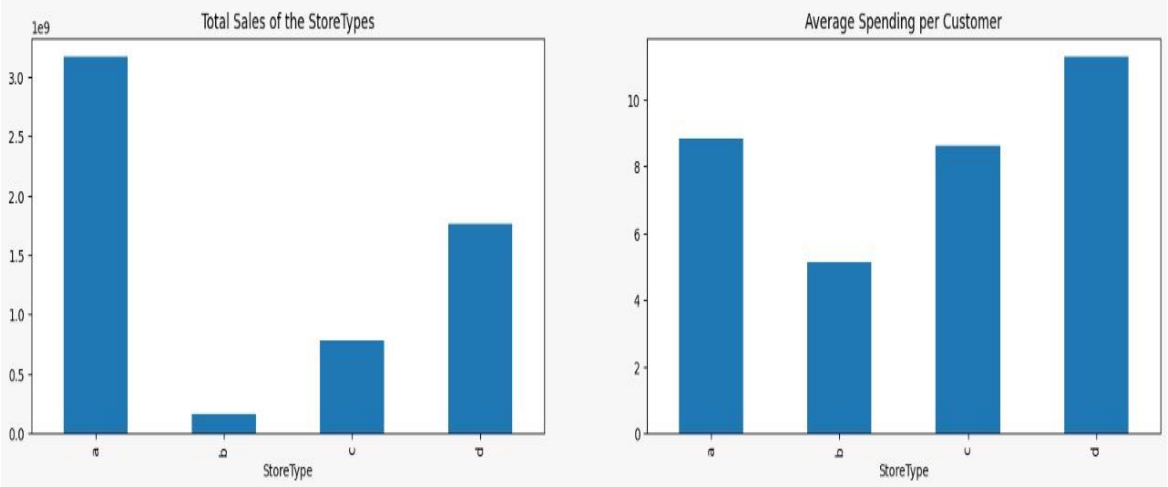


We have used Scatterplot to visualize the trend of Customers and Competition Distance with Sales. We can see a linear relationship between the two.

Plot used: Scatterplot

Data Visualization

Store Type and Sales & Average Spending per Customer

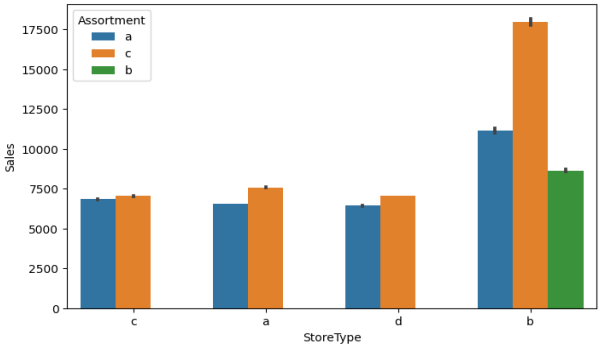


- Store ‘a’ has maximum customers.
- Store ‘d’ has maximum average spendings per customers.

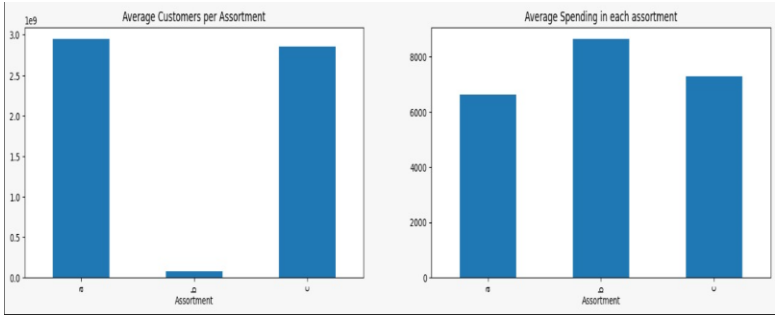
Plot used: Histogram

Data Visualization

Store Type and Sales



Assortment and Sales



To visualize the assortment-wise sales in each Store Type, we used bar plots. We can see that only Store Type "B" has all three assortment levels a, b, and c.

Assortment and Sales

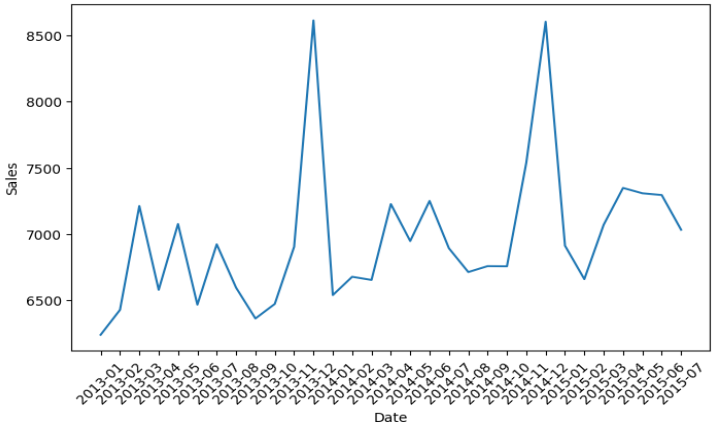
The histograms shows Average Customers per Assortment and Average Spending in each assortment. We can see that:

- Assortment a and c has maximum customers.
- In terms of average spending per customers assortment b has maximum footfalls.

Plot used: Barplot and Histogram

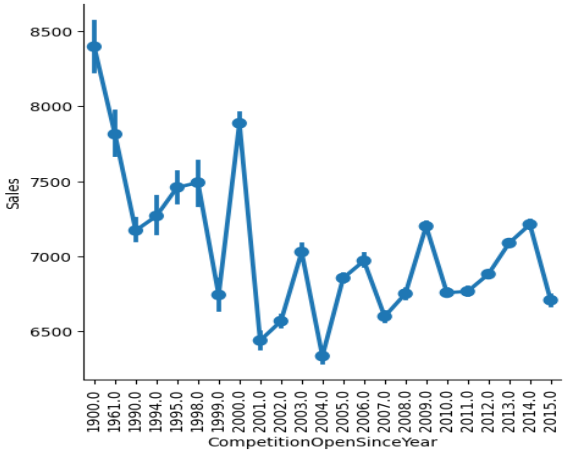
Data Visualization

Date and Sales & Competition Open Since Year and Sales



Monthly Sales

Competition Open Since Year and Sales

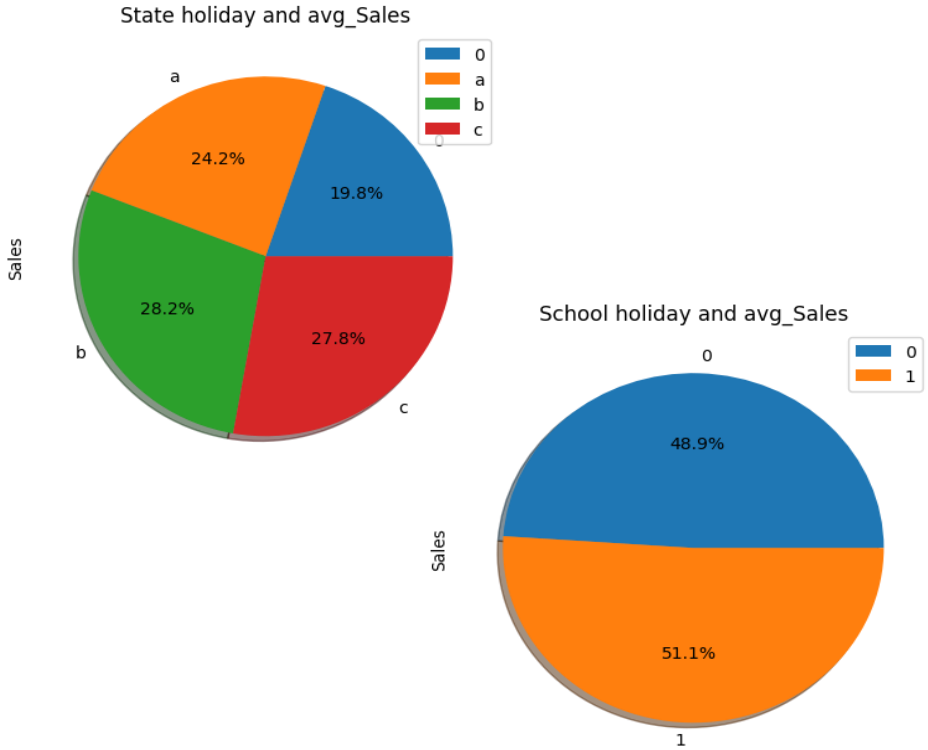


We have used Line plots to visualize the Trend of Sales across months and Trend of Sales due to Competition Open Since Year. Here, it is clear that October 2013 and 2014 had the highest sales, demonstrating that October is the best month for sales.

Plot used: Lineplots

Data Visualization

State Holiday & School Holiday with Average Sales

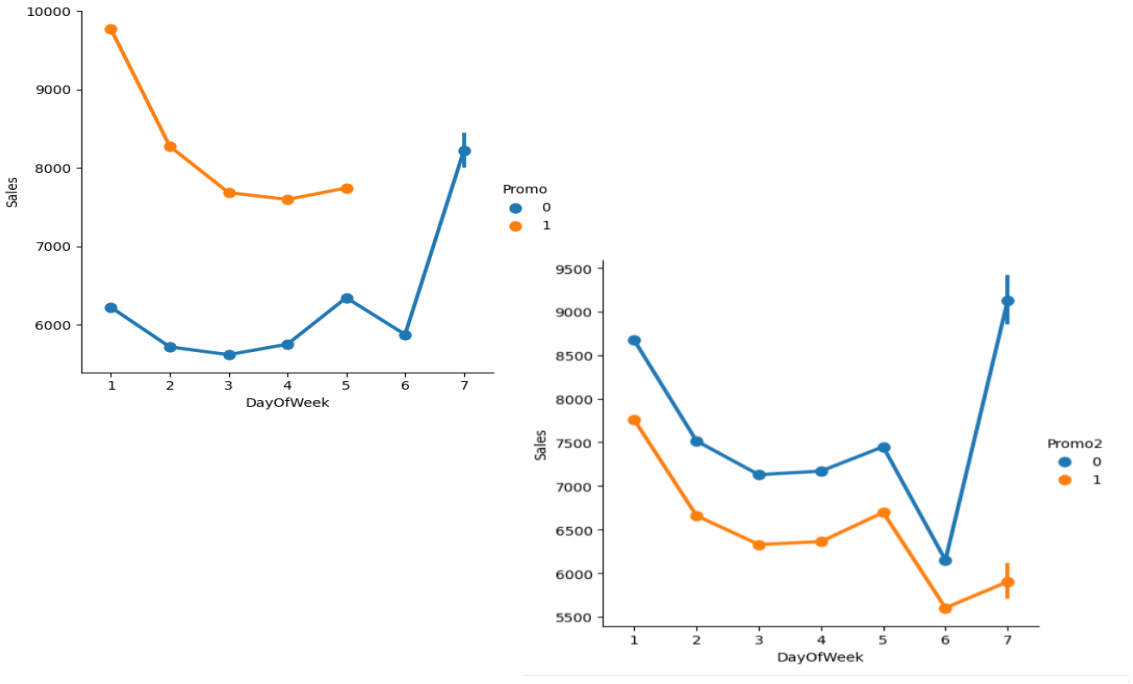


It is clearly evident that on holidays average sales are better as compared to non holiday.

Plot used: Pie Charts

Data Visualization

Promo & Promo2 with Sales

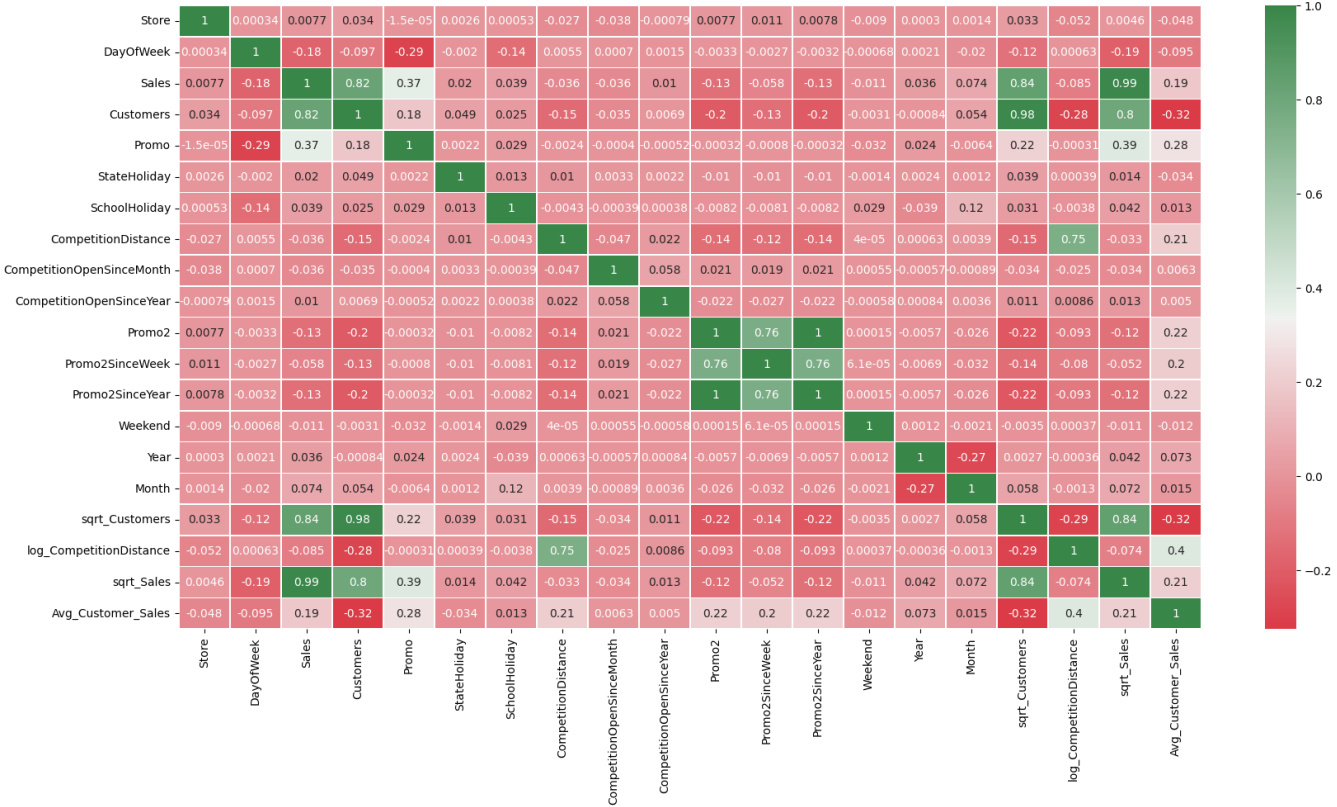


Here we have used Line plots to see the relationship. It can be seen that there's no promotion over the weekend however the sales are very high. Sales will undoubtedly bounce if promotional deals are made available on the weekends.

Plot used: Line Plots

Data Visualization

Multivariate Analysis



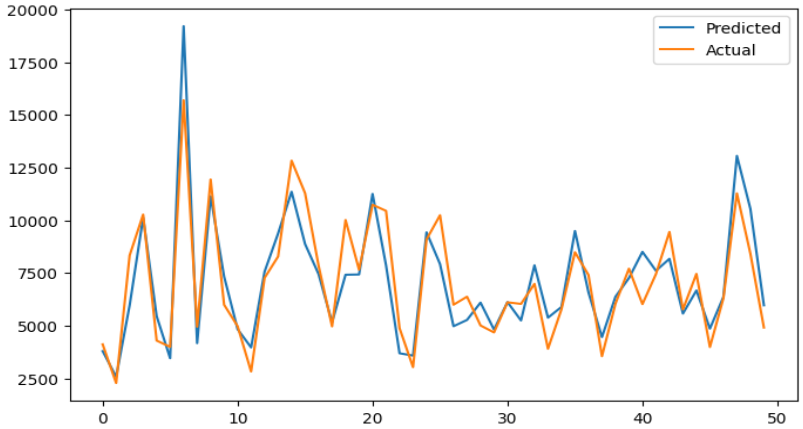
To gain a sense of the correlation between the independent variables and the target variable, we used Heatmap. This will also be used in Feature Engineering. We can see Promo2 is highly correlated with promo2sinceYear and promo2sinceWeek.

Plot used: Heatmap

Implementation of ML Models

- Linear Regression
- LASSO Regression
- RIDGE Regression
- Decision Tree
- Random Forest

Linear Regression

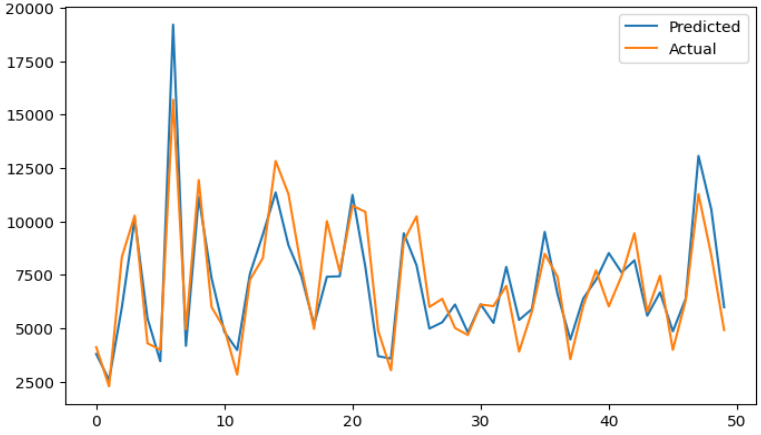


Regresion Model Score : 0.8632730979707379
Out of Sample Test Score : 0.8620825434853236

Training RMSE : 6.460913572181647
Testing RMSE : 6.474678884573082

Training MAPE : 6.2395040946099485
Testing MAPE : 6.264130235846424

LASSO Regression

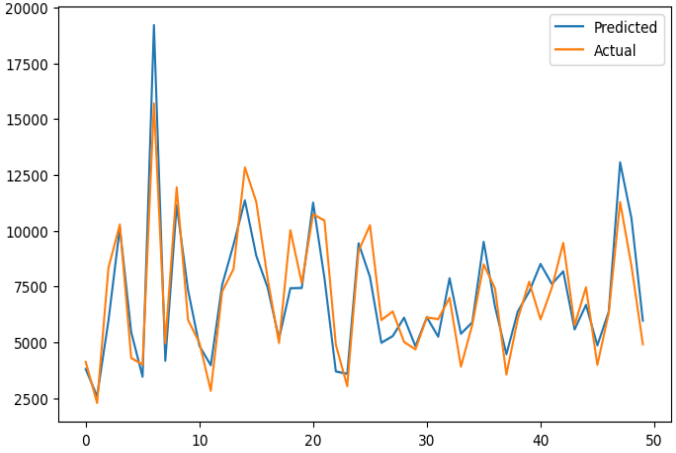


Regresion Model Score : 0.8632790801198456
Out of Sample Test Score : 0.8620839187158411

Training RMSE : 6.460772229949226
Testing RMSE : 6.474646603676903

Training MAPE : 6.240653544026185
Testing MAPE : 6.265303166288976

RIDGE Regression

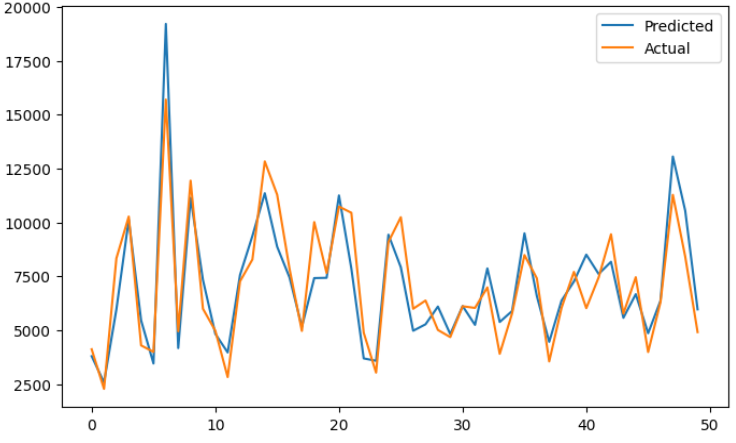


Regresion Model Score : 0.8632790800960524
Out of Sample Test Score : 0.8620839324953599

Training RMSE : 6.460913572181647
Testing RMSE : 6.474678884573082

Training MAPE : 6.2395040946099485
Testing MAPE : 6.264130235846424

Decision Tree

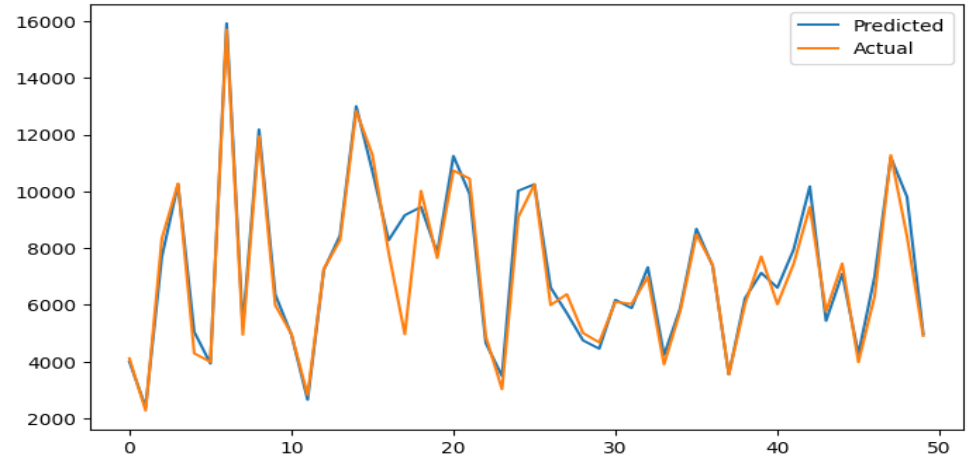


Regresion Model Score : 0.9637035071087259
Out of Sample Test Score : 0.9549910905462698

Training RMSE : 6.460913572181647
Testing RMSE : 6.474678884573082

Training MAPE : 6.2395040946099485
Testing MAPE : 6.264130235846424

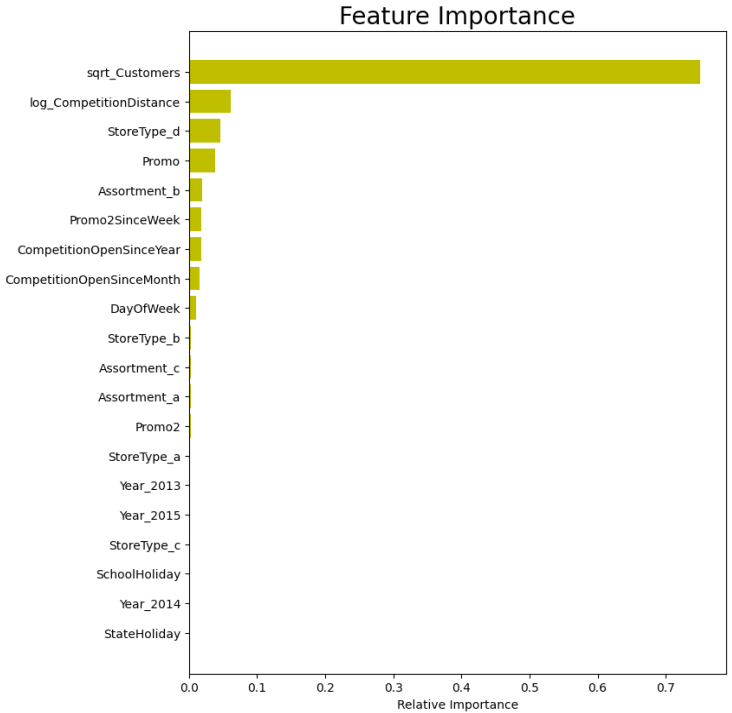
Random Forest



Regresion Model Score : 0.9954749309285421
Out of Sample Test Score : 0.9723550814287851

Training RMSE : 1.175383395149589
Testing RMSE : 2.8987867804031278

Training MAPE : 1.0351845909717141
Testing MAPE : 2.631437860623957



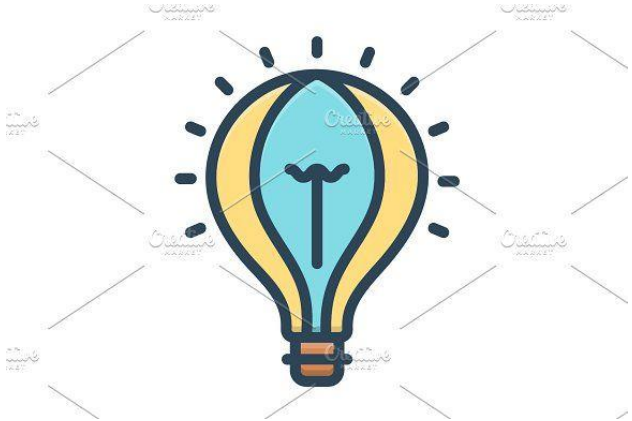
Evaluation Metrics

Training Score: Accuracy of the Model on Training Dataset
Testing Score: Accuracy of the Model on Testing Dataset
RMSE: Root Mean Squared Error
MAPE: Mean Absolute Percentage Error

Metrics/ Models	Linear Regression	LASSO Regression	RIDGE Regression	Decision Tree	Random Forest
Training Score	0.863	0.863	0.863	0.96	0.99
Test Score	0.862	0.862	0.862	0.95	0.97
Training RMSE	6.46	6.46	6.46	71.06	1.17
Testing RMSE	6.47	6.47	6.47	70.96	2.90
Training MAPE	6.23	6.23	6.24	83.74	1.03
Testing MAPE	6.26	6.26	6.26	83.73	2.63

Conclusion

With 99% Training Accuracy and 97% Testing Accuracy, Random Forest has proven to be the most efficient model out of the algorithms used in our model, including Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, and Random Forest.



Recommendations

- As per an observed pattern, shops that are far from their competitors are less likely to experience sales than shops that are close to one another. As a result, it is recommended that stores be opened close to those of its competitors.
- Since October had the highest sales across all three years, more offers and promotions should be made available during that time period to capitalize on the advantage.
- It is discovered that customers in Assortment B spend a lot of money, irrespective of there small numbers. If some targeted advertisements are pushed to the wealthy customers, sales in this assortment could rise.
- Even if there is no promotion running over the weekend, there are still significant sales. It can be recommended to offer various incentives in the weekend to boost sales.
- Only Store Type B carries all three of the offered assortments (a, b, and c). Stores for c type assortment levels are lacking for other store types, so efforts should be made to fill these gaps in order to increase revenue.

Thank You!!

The End of a Story.....
The Beginning of Many