

Capstone Project - 2

Retail Sales Prediction By



"Suraj kad"

suraj.kad.90@gmail.com

Data Science Trainees

Content:

- Problem Statement
- Data Summary
- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- Model Implementation
 - Linear Regression
 - Lasso Regression
 - Decision Tree Regression
- Conclusion and Recommendations



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.

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.



Data Summary:

We have two datasets. Rossman store data is for years 2013, 2014 and 2015 with 10,17,209 observations on 9 variables. Stores data with 1115 observations on 10 variables. Some important features are: -

- **Id** - an Id that represents a (Store, Date) duple within the set
- **Store** - a unique Id for each store
- **Sales** - the turnover for any given day (Dependent Variable)
- **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
- **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. An assortment strategy in retailing involves the number and type of products that stores display for purchase by consumers.
- **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.



Approach

The following approach was followed in the completion of the project:

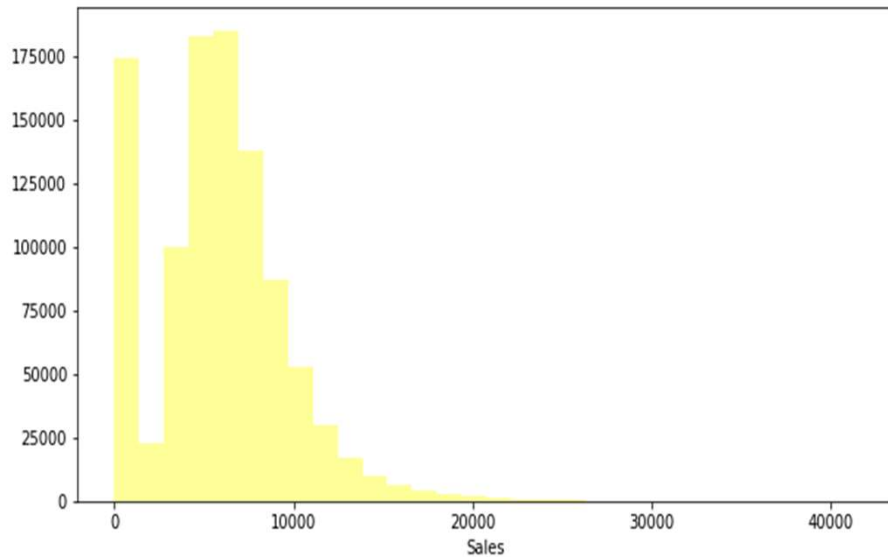
- Business Problem
- Data Collection and Preprocessing
 - Data Cleaning
 - Missing Data Handling
 - Merging the Datasets
- Exploratory Data Analysis
- Data Manipulation
 - Feature Engineering
 - Outlier Detection and Treatment
 - Feature Scaling
 - Categorical Data Encoding
- Modeling
 - Train Test Split
 - Linear Regression
 - Lasso Regression
 - Decision Tree Regression
 - Random Forest
- Model Performance and Evaluation
- Conclusion and Recommendations

Exploratory Data Analysis

Hypotheses

Just by observing the head of the dataset and understanding the features involved in it, the following hypotheses could be framed:

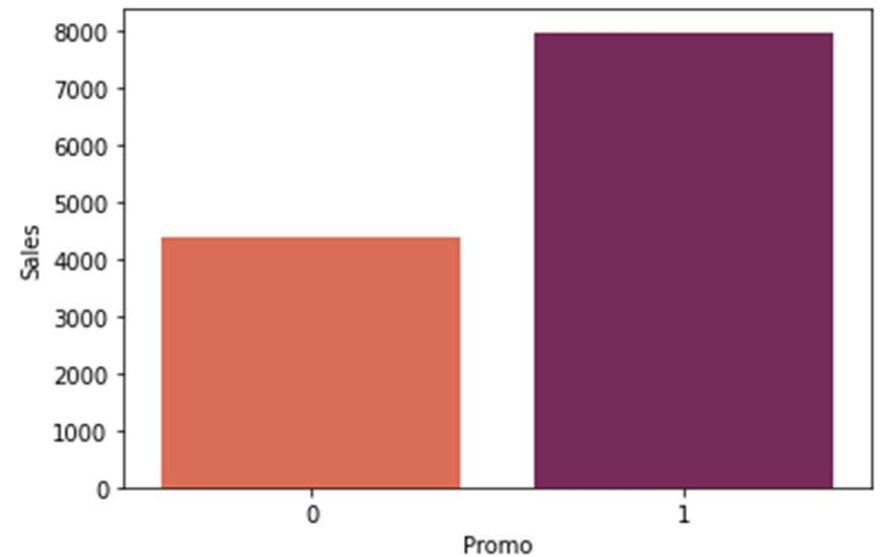
- There's a feature called "DayOfWeek" with the values 1-7 denoting each day of the week. There would be a week off probably Sunday when the stores would be closed and we would get low overall sales.
- Customers would have a positive correlation with Sales.
- The Store type and Assortment strategy involved would be having a certain effect on sales as well. Some premium high quality products would fetch more revenue.
- Promotion should be having a positive correlation with Sales.
- Some stores are closed due to refurbishment, those would generate 0 revenue for that time period.
- There would be some seasonality involved in the sales pattern, probably before holidays sales would be high.



Sales are normally distributed with slightly right tail skewed.

Histogram Representation of Sales. Here 0 is showing because most of the time store was closed

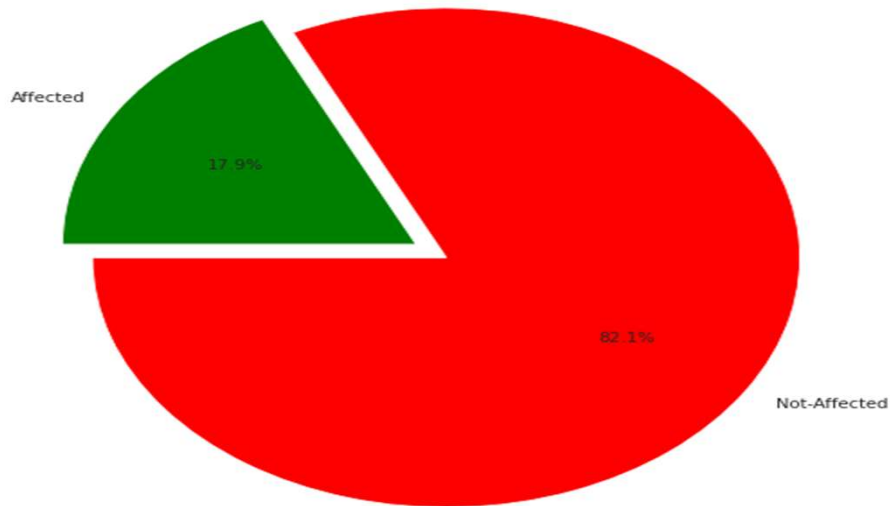
- Impact of Promo on sales



Sales Are nearly doubled High When Promotion is Running.

- we can understand the school holiday are affected on sales or not

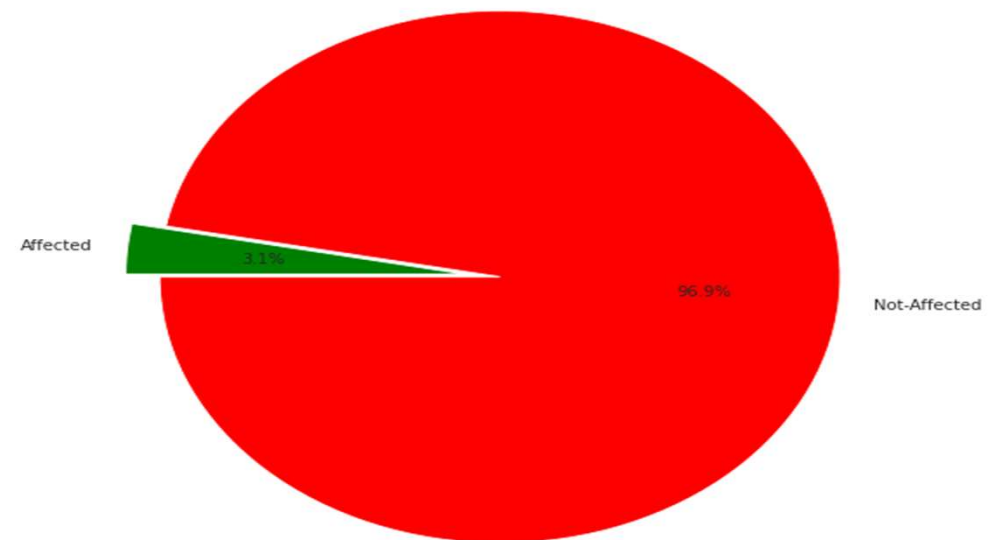
Sales Affected by Schoolholiday or Not ?



"As we can see in the Pie chart... Sales affected by School Holiday is 18% and Mainly Sales aren't affected by School Holiday"

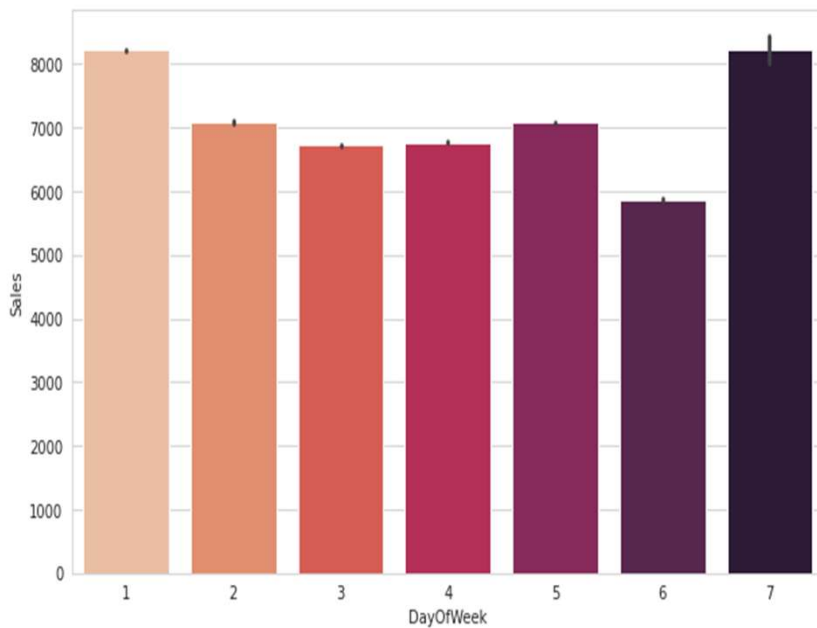
- Understand can state holiday affected on the sales?

Sales Affected by State holiday or Not ?



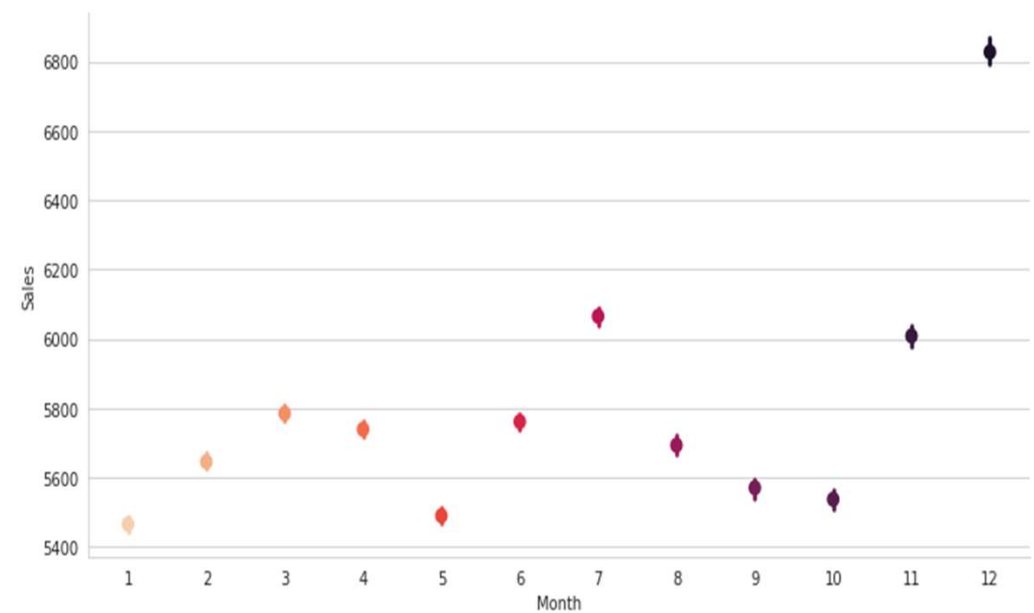
As we can see in the Pie chart Sales affected by State Holiday is only 3% means Sales aren't affected by State Holiday. As Sales isn't much affected by State Holiday so I'm removing this column

- Day Wise trends in Sales



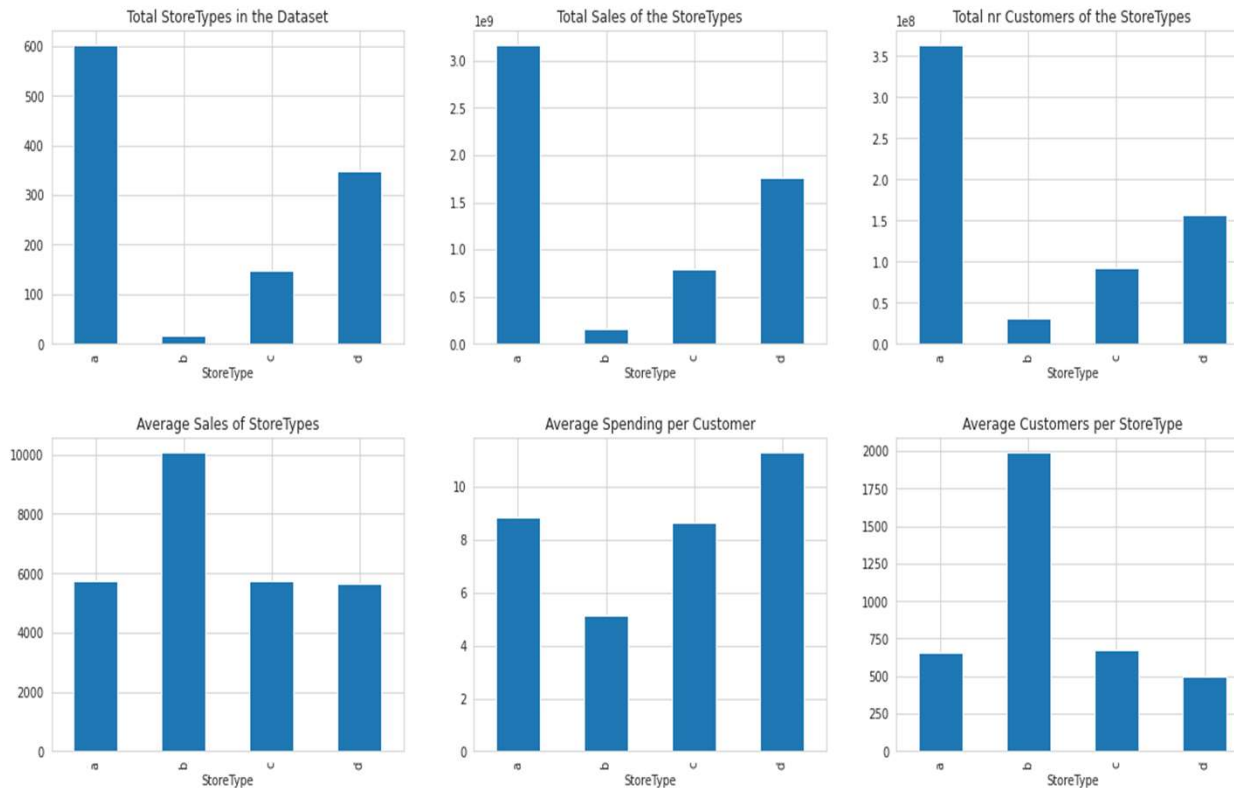
From plot it can be seen that most of the sales have been on 1st and last day of week.

- Monthly trends in Sales



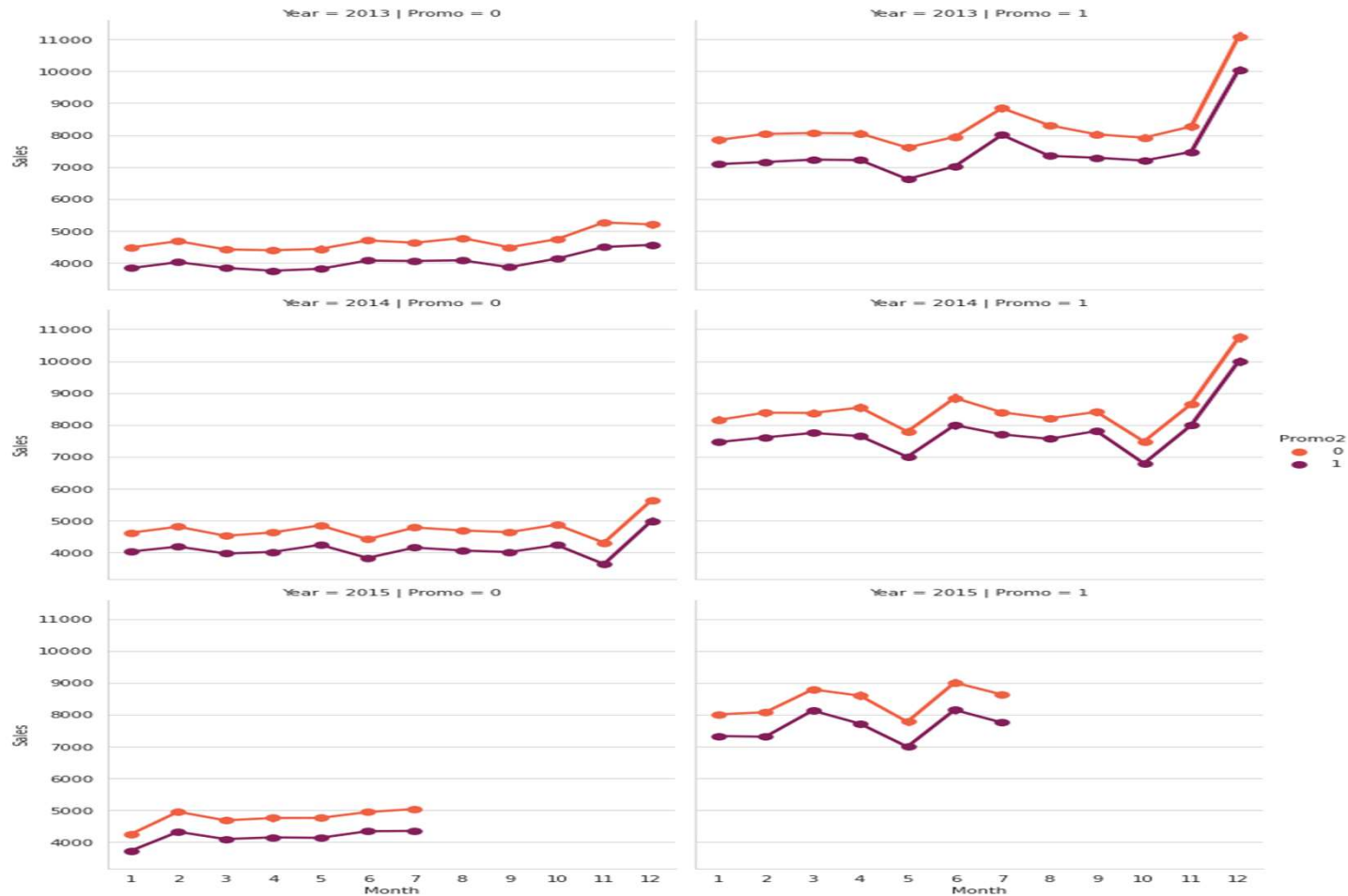
As we can see that in the month of November and specially in December sales are increasing rapidly every year on Christmas.

Store Types and average sales/customer/spending relation

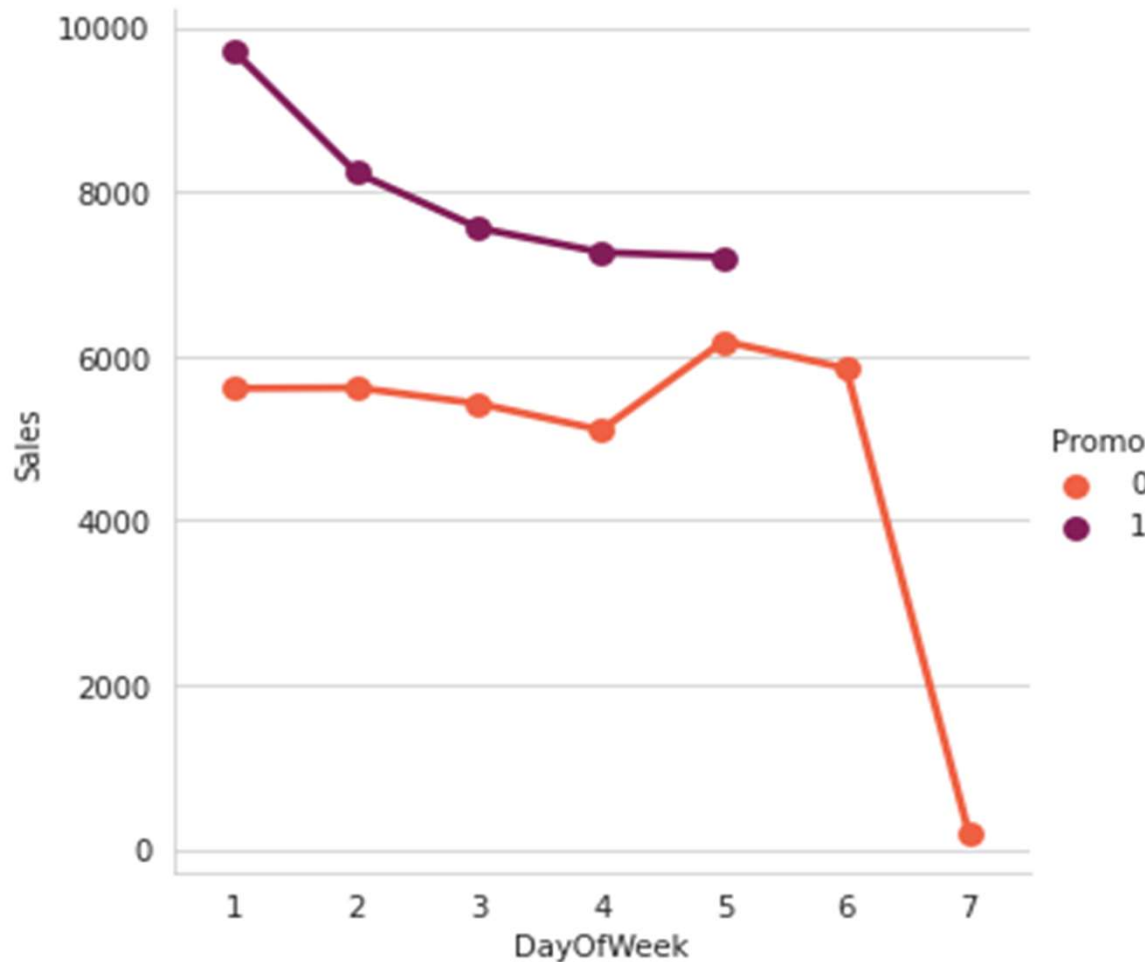


As we can see from the graphs, the Store Type A has the most stores, sales and customers. However, the Store Type D has the best averages pending per customers. Store Type B, with only 17 stores has the most average c

- Understand can promotion affected on Sales

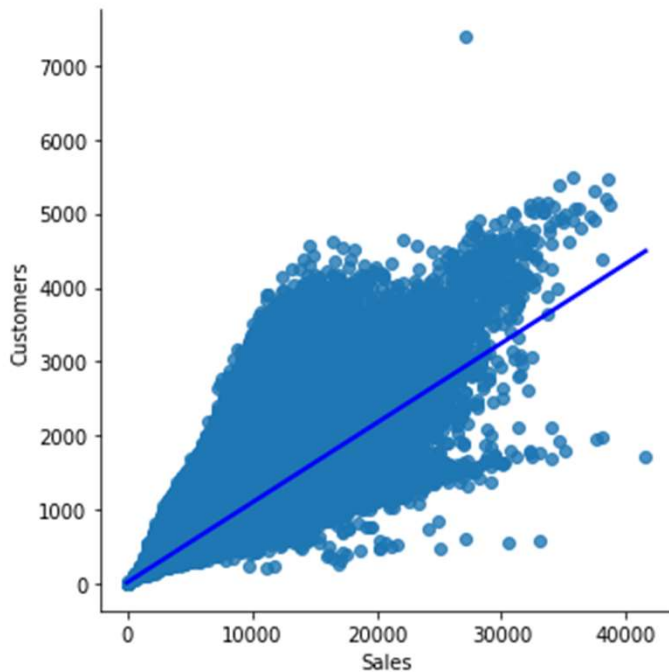


So, of course, if the stores are having promotion the sells are higher. Overall the store promotions selling's are also higher than the seasonality promotions (Promo2). However I can't see no yearly trend.



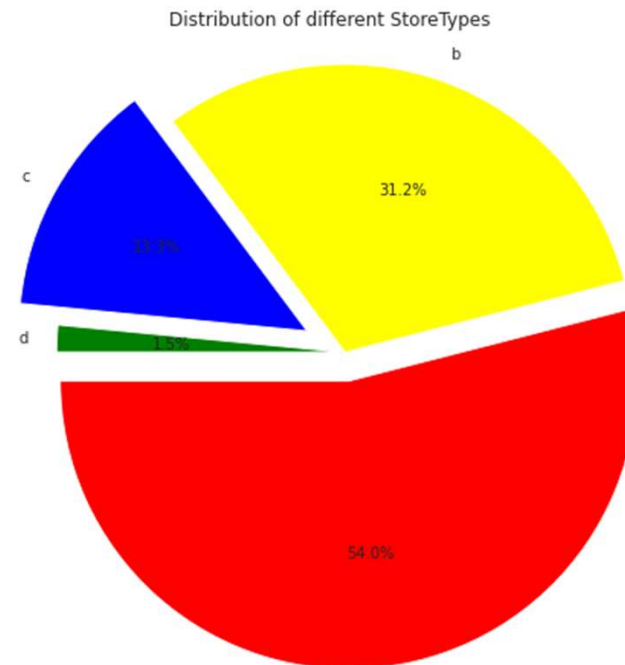
As We can see that when the promotion is running Sales are high.
So, no promotion in the weekend.
However, the sales are very high, if the stores have promotion. The Sales are going crazy on Sunday. No wonder.

- Sales VS Customer



As we can see from the graphs, the Sales are correlated in a customers.

- Distributions of different store type



This pie chart show in distribution of store name are a,b,c and d respectively. So a store distribution is 54% ,second b store distribution is app. 32% , c store is 23% and last store d is a 1.5%

EDA Conclusion:

- Sales are highly correlated to customers.
- Stores opened on 'State Holiday' makes a good amount of sales.
- There is no such significant difference in sales on 'School Holidays'.
- Even though store type 'b' has very a smaller number of stores but these are outperforming other store types in terms of sales and avg customers.
- Sales are consistent for the second quarter of the year but it starts increasing in the last quarter.

Modeling:

Factors affecting in choosing the model:

Determining which algorithm to use depends on many factors like the problem statement and the kind of output you want, type and size of the data, the available computational time, number of features, and observations in the data, to name a few.

The dataset used in this analysis has:

- A multivariate time series relation with sales and hence a linear relationship cannot be assumed in this analysis. This kind of dataset has patterns such as peak days, festive seasons etc. which would most likely be considered as outliers in simple linear regression.
- Having X columns with 30% continuous and 70% categorical features. Business prefers the model to be interpretable in nature and decision based algorithms work better with categorical data.

- Linear Regression

```

✓ [215] print("Linear_R_model_conclusion")

print("Linear_R_model_Score-->>>", LinearRegression_train_score)

print("Linear_R_model_Sample_test score-->>>", LinearRegression_test_score)

print("Root MEan Squar Error ")
print("Train_RMSE --->>>", Train_RMSE , "Test_RMSE --->>>", Test_RMSE)

print("Mean Absolute percentage error (MAPE) ")
print("Train_MAPE --->>>", Training_MAPE , "Test_MAPE --->>>", Testing_MAPE)

```

```

Linear_R_model_conclusion
Linear_R_model_Score-->>> 0.9407624885169805
Linear_R_model_Sample_test score-->>> 0.9405188463467887
Root MEan Squar Error
Train_RMSE --->>> 570439.8485694885 Test_RMSE --->>> 573373.5891347087
Mean Absolute percentage error (MAPE)
Train_MAPE --->>> 0.06843341999997476 Test_MAPE --->>> 0.06868939986477175

```

- Lasso Regression

```

✓ [223] print("Lasso_R_model_conclusion")

print("Lasso_R_model_Score-->>>", LassoR_train_score)

print("Linear_R_model_Sample_test score-->>>", LassoR_test_score)

print("Root MEan Squar Error ")
print("Train_RMSE --->>>", Train_RMSE_LR , "Test_RMSE --->>>", Test_RMSE_LR)

print("Mean Absolute percentage error (MAPE) ")
print("Train_MAPE --->>>", Training_MAPE_LR , "Test_MAPE --->>>", Testing_MAPE_LR)

```

```

Lasso_R_model_conclusion
Lasso_R_model_Score-->>> 0.7521725153605543
Linear_R_model_Sample_test score-->>> 0.7514805019619203
Root MEan Squar Error
Train_RMSE --->>> 2386505.936354313 Test_RMSE --->>> 2395624.627437886
Mean Absolute percentage error (MAPE)
Train_MAPE --->>> 0.1916114432081867 Test_MAPE --->>> 0.19134833760235753

```

- **Decision Tree**

```
✓ [236] print("Decision Tree Regression_model_conclusion")
15
print("Decision Tree Regression_model_Score-->>>", DTR_train_score)

print("Decision Tree Regression_model_Sample_test score-->>>", DTR_test_score)

print("Root MEan Squar Error ")
print("Train_Decision Tree Regression --->>> ", Train_RMSE_Tree , "Test_RMSE --->>> ", Test_RMSE_Tree)

print("Mean Absolute percentage error (MAPE) ")
print("Train_Decision Tree Regression --->>> ", Training_MAPE_Tree , "Test_MAPE --->>> ", Testing_MAPE_Tree)
```

```
Decision Tree Regression_model_conclusion
Decision Tree Regression_model_Score-->>> 1.0
Decision Tree Regression_model_Sample_test score-->>> 0.9995839250787935
Root MEan Squar Error
Train_Decision Tree Regression --->>> 0.0 Test_RMSE --->>> 4010.7892377478265
Mean Absolute percentage error (MAPE)
Train_Decision Tree Regression --->>> 0.0 Test_MAPE --->>> 0.0022376776007860684
```

- A baseline is a simple model that provides reasonable results on a task and does not require much expertise and time to build. It is well established that there is seasonality involved and no linear relationship is possible to fit. For these kinds of datasets tree based machine learning algorithms are used which are robust to outlier effect which can handle non-linear data sets effectively.
- The results show that a simple decision tree is performing pretty well on the validation set but it has completely overfitted the train set. It's better to have a much more generalized model for future data points.

- **Random Forest**

```

✓ 0s ▶ print("Random Forest_model_conclusion")

print("Random Forest_model_Score-->>>", RDF_train_score)

print("Random Forest_model_Sample_test score-->>>", RDF_test_score)

print("Root MEan Squar Error ")
print("Train_Random Forest --->>>" ,Train_RMSE_rdfreg , "Test_Random Forest RMSE --->>>", Test_RMSE_rdfreg)

print("Mean Absolute percentage error (MAPE) ")
print("Train_Random Forest --->>>" ,Training_MAPE_rdfreg , "Test_Random Forest MAPE --->>>", Testing_MAPE_rdfreg)

```

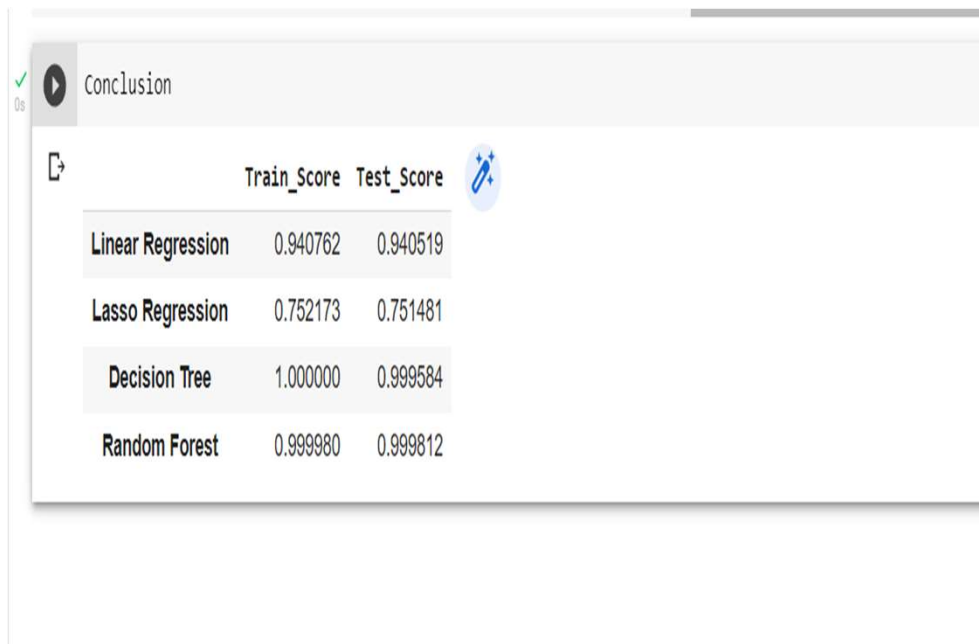
```

❏ Random Forest_model_conclusion
Random Forest_model_Score-->>> 0.9999795149579371
Random Forest_model_Sample_test score-->>> 0.9998120687376748
Root MEan Squar Error
Train_Random Forest --->>> 197.2649424286102 Test_Random Forest RMSE --->>> 1811.5792275697581
Mean Absolute percentage error (MAPE)
Train_Random Forest --->>> 0.000307406587990983 Test_Random Forest MAPE --->>> 0.0007316462715106348

```

- Random forests are an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time. For regression tasks, the output of the random forest is the average of the results given by most trees.
- To prevent overfitting, we built random forest model. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Model Evaluation



Conclusion

| | Train_Score | Test_Score |
|-------------------|-------------|------------|
| Linear Regression | 0.940762 | 0.940519 |
| Lasso Regression | 0.752173 | 0.751481 |
| Decision Tree | 1.000000 | 0.999584 |
| Random Forest | 0.999980 | 0.999812 |

Predictions from random forest model are very close to actual values in our X dataset as we have good score. The figure shows actual values, predicted & the difference between them respectively.

Since this is Sales prediction MAE is a good metric.

We're getting Mean Absolute Error ~ \$ 197.23

And MAPE of 0.007%

Conclusion:

Our model shows that Customers, Competition distance, Store type are some of the most important features in our sales prediction.

We need to focus on these aspects to maximize our profits for the next 6 weeks.

AI



**Thank
You!!!**

AImaBetter

AImaBetter