Capstone Project - 2

Rossmann Sales Prediction

Project by- Potdar Vinayak

Contents

- Abstract
- Problem Statement
- Understanding Datasets
- Null Value Analysis
- Correlation and Multi-Collinearity Check
- EDA
- Feature Engineering and Selection
- Utility Functions
- Model Fittings and Feature Importance
- Conclusion

Abstract

In this experiment, we built a predictive model for a continuous data. The task was to build a predictive model, which would help in predicting the sales of a drug store chain.

We started by understanding the data and performed EDA which was then followed by ML model fittings.

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.

Understanding the Datasets

We have been provided with two datasets, the first one focuses on sales and customer data whereas the second one focuses on the supplementary part of the dataset.

1. Rossmann Store Data- This data has all the records of sales. It houses the additional data about the number of customers and promotional offer availability followed by the StateHoliday and SchoolHoliday features which represent an effect on sales.

Understanding the Datasets

2. Store Data- This data consists of all the records regarding each store. This houses some critical data that can be considered as external effects on sales. This data houses features which describe competition details, store types, assortment level of every particular store and some more information about another promotional offer. This data acts as a supplementary dataset on the first, which helps us understand the behaviour of each record in the first dataset.

Null Value Analysis and Treatment

The Rossmann Store Data has no null values present in it but the supplementary dataset i.e. Store Data has five features with major missing values ranging from 31% to 48% and one feature with 3 missing values.

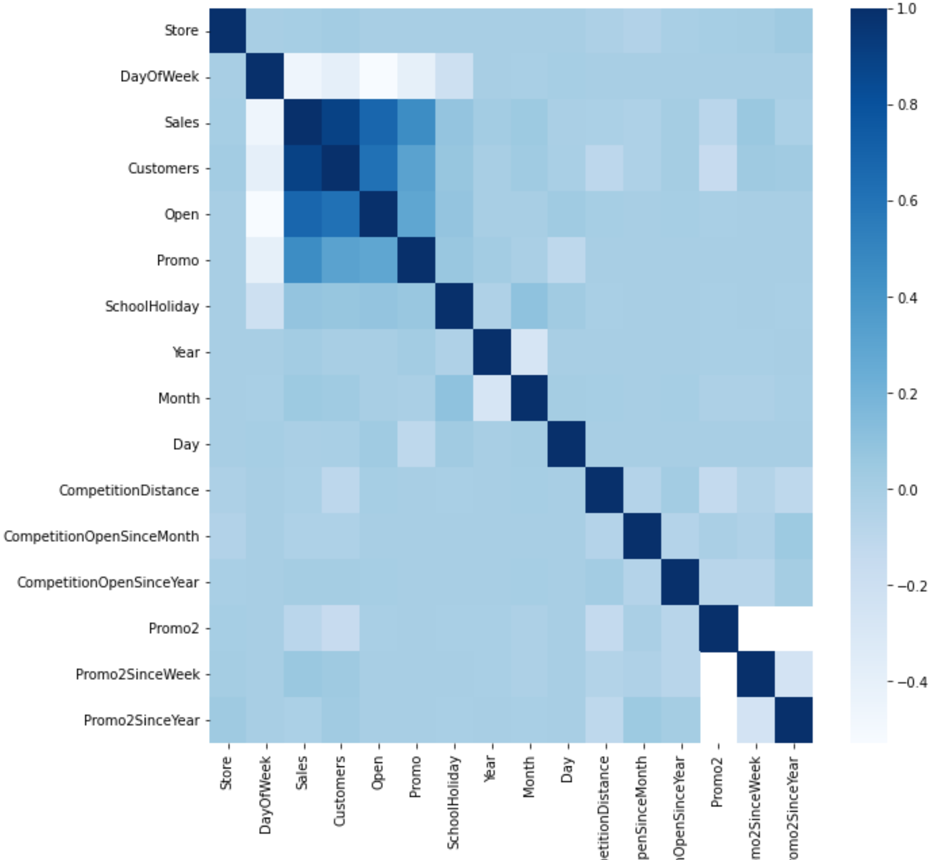
Until we understand the relation and relevance of all the features, dropping them would not be the right decision. Here we just treated the CompetetionDistance feature with 3 missing values, where we replaced the missing values with the median value. We used median value as it is not affected by the rest of data.

We dropped some features in the feature selection phase.

Correlation and Multi-Collinearity Check

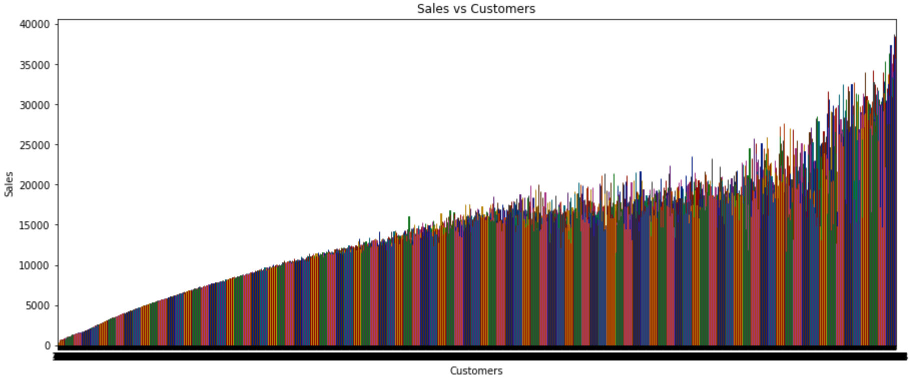
We merged both the datasets and formed a correlation matrix to understand the relation between the features.

This matrix does not show any features with high collinearity, but we can see the correlation of some features with Sales pretty clearly.



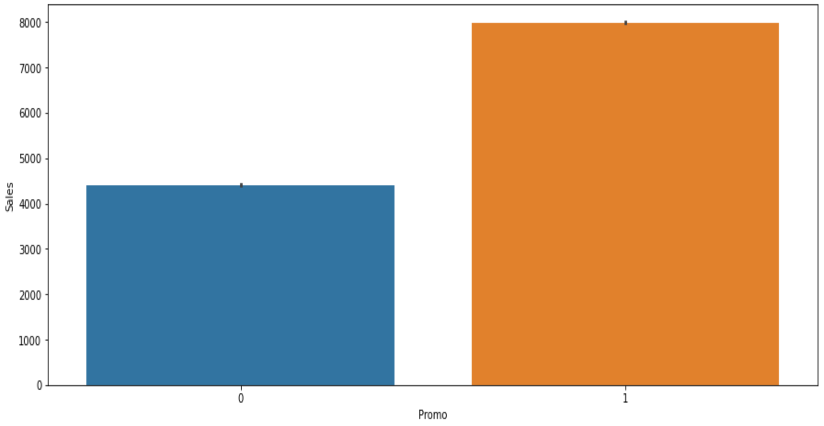
EDA

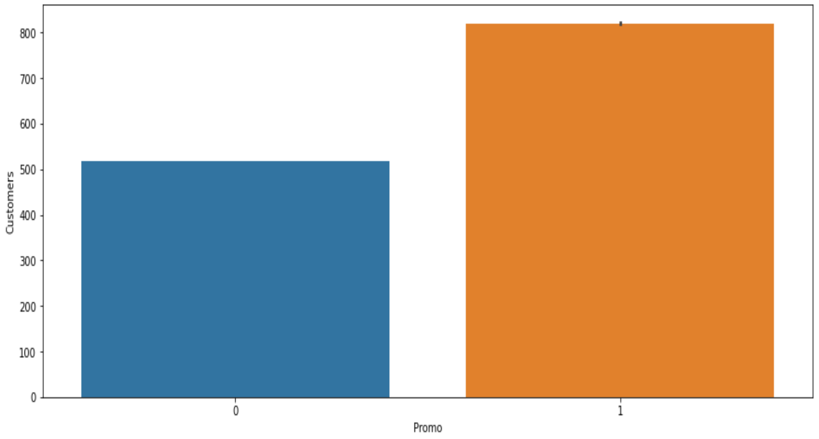
We started with getting the relation between Sales and Customers , as it showed the most positive relation in the correlation matrix. This shows a fairly linear relation in both, indicating that there majority of the customers coming in the store, tend to buy something. Due to this reason, we analyzed other features with no. of customers also.



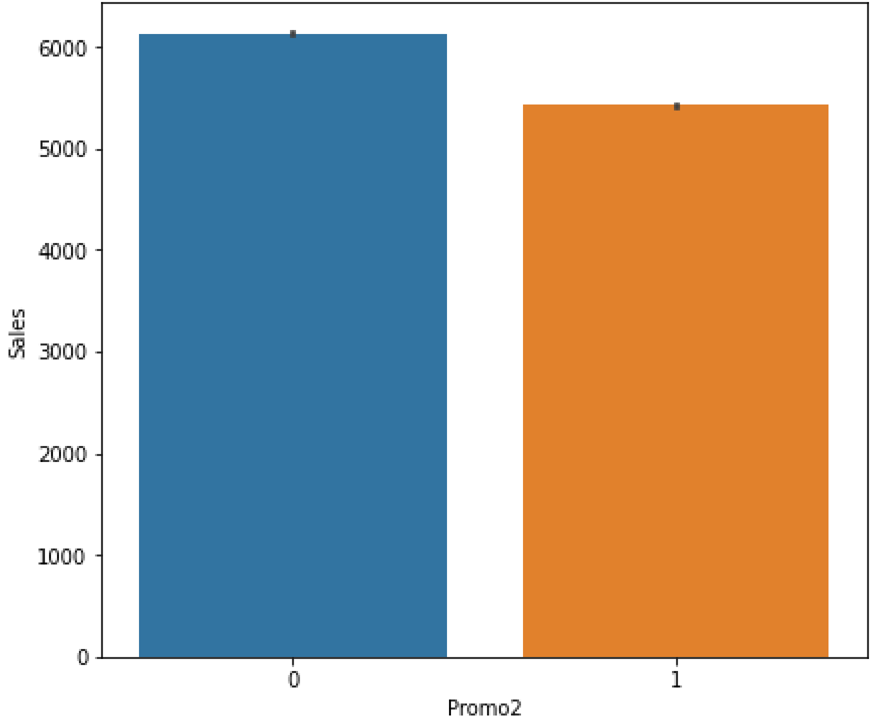
EDA

A littile distortion in the sales can be seen in the higher region of previous graph. To verify if it was due to price of the products, we did some hypothesis testing. We compared the effects of promotional offers on sales and figured out; it has an effect on sales but not significant.

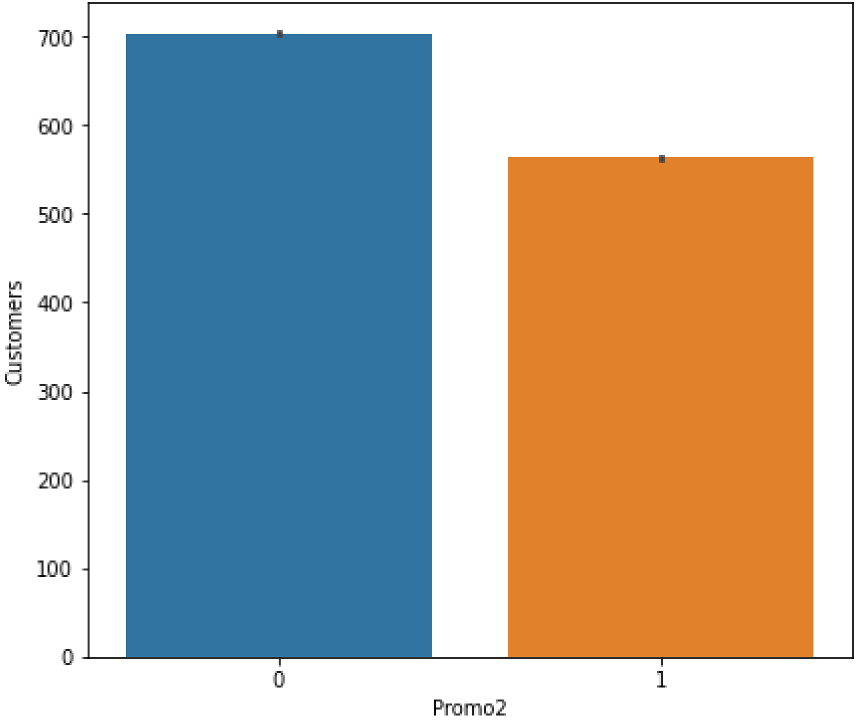




EDA



Sales
6000
5000
4000
3000
2000
1000
0
0
1
Promo2

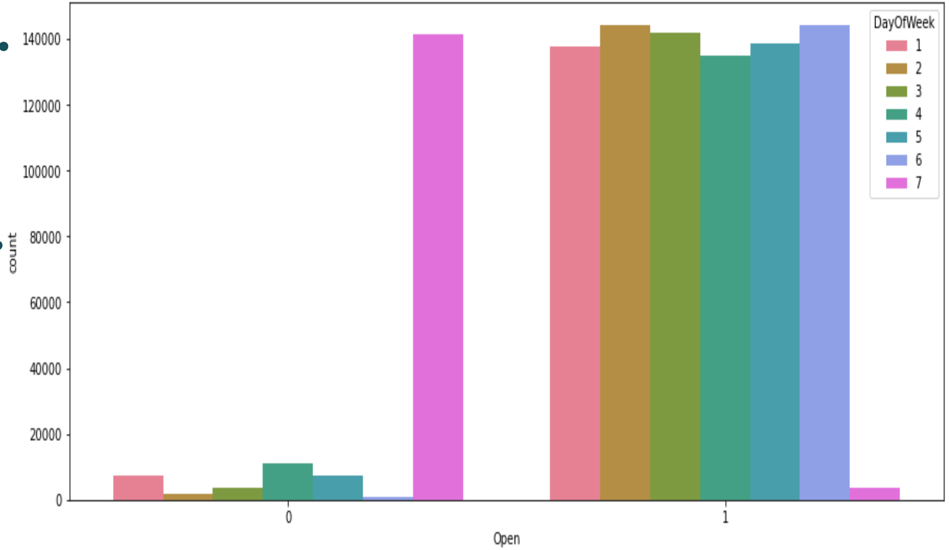


Customers
700
600
500
400
300
200
100
0
0
1
Promo2

EDA

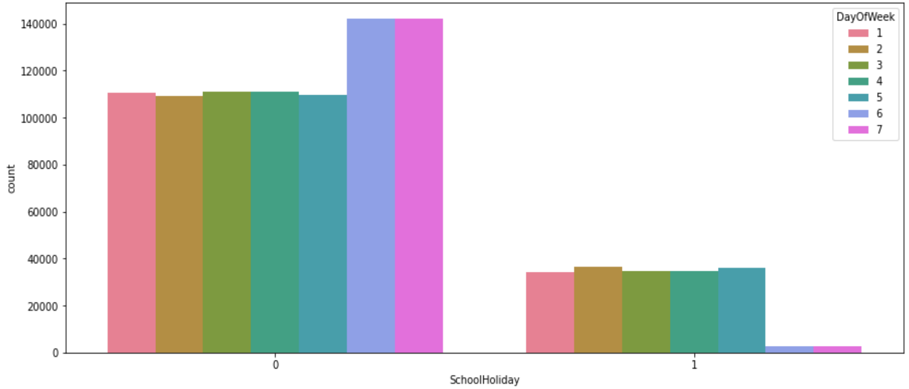
We proceeded with understanding the operational days of the stores. Here we can see that the stores are mostly active during the weekdays and closed on sundays.

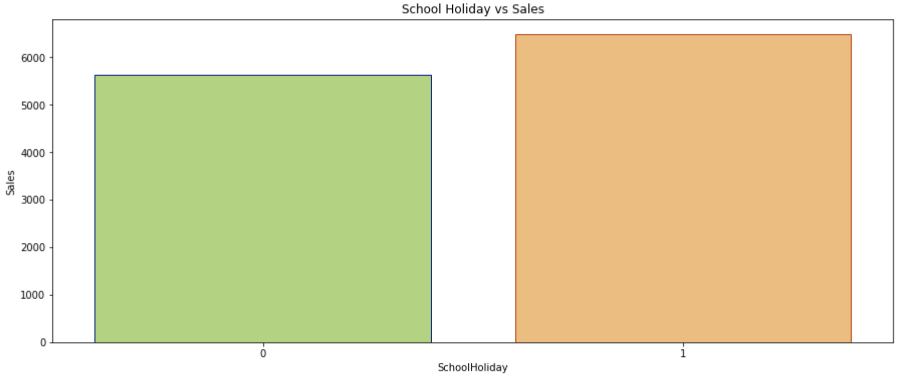
We see some exceptions, but these can be considered as minority.



EDA

Here, we examined the affects of school holiday on the sales of the stores. The schools are not operational on the weekends. We see a higher sales count when the schools are closed i.e. mostly on weekends. The direct reason cannot be deduced, but we can assume that people are refilling their prescriptions over the weekend, which adds to the regular sales count.

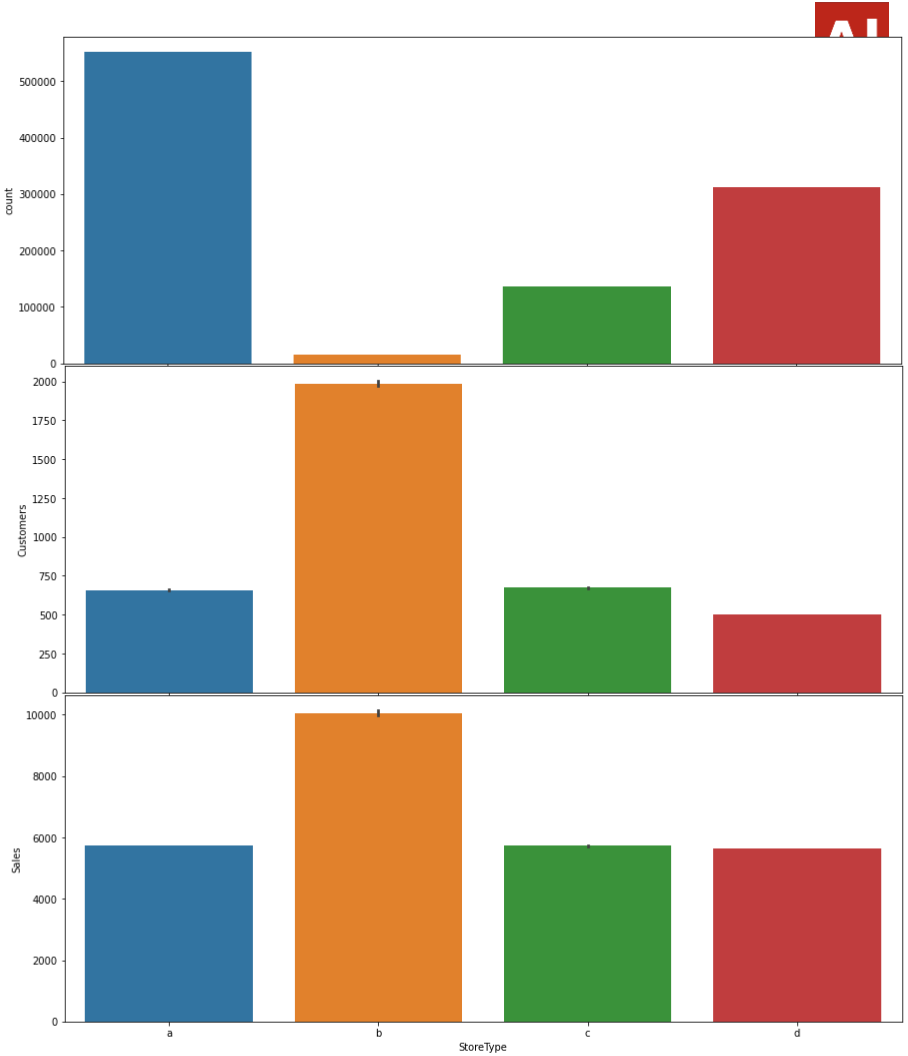




EDA

StoreType vs Customers Vs Sales:

Store type B has the least no. of stores but highest no. of customers and sales. Store Type A has the most no. of stores but fairly consistent customers and sales. Stores C and D show good customer to sales ratio when compared with the number of stores.



EDA

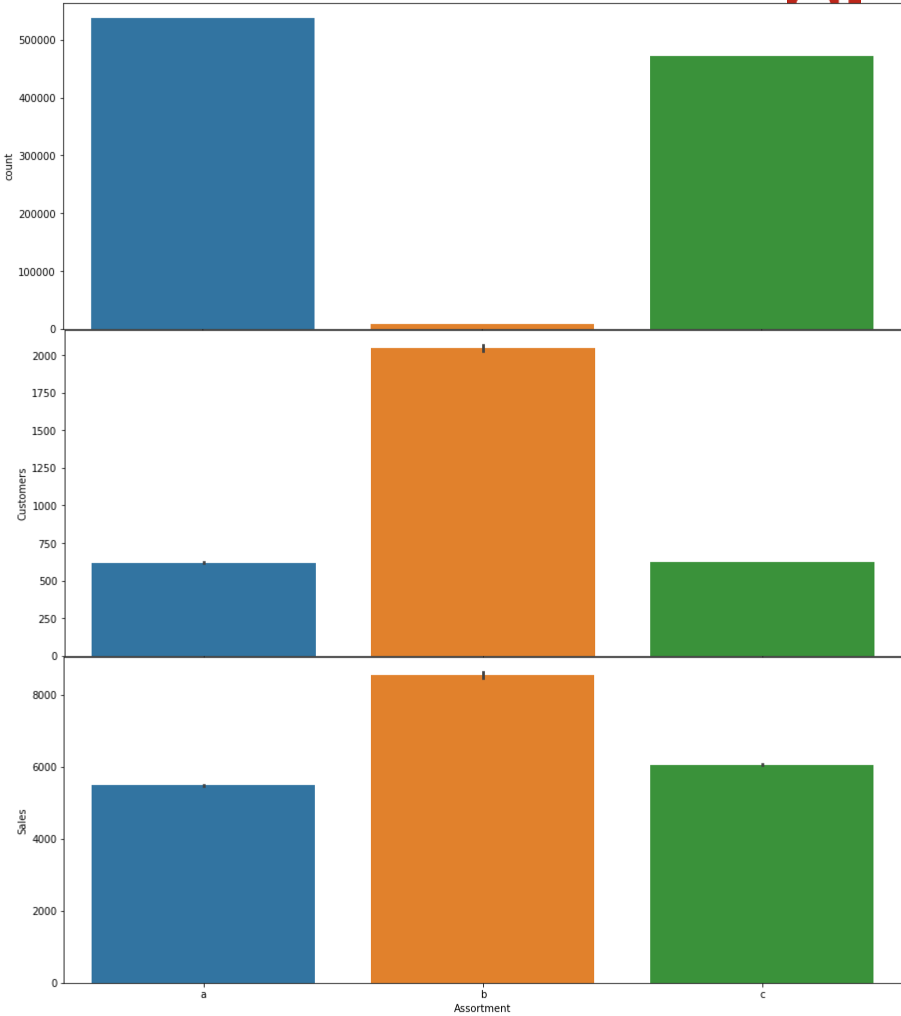
Assortment vs Customers Vs Sales:

There are 3 assortments Basic, Extra and Extended represented by a, b and c respectively.

Assortment B has the least no. of stores but the highest customers and sales.

Assortment A and C most no. of stores with lesser sales and customers than B.

Interestingly A and C have a better Sales to customer ratio than B.



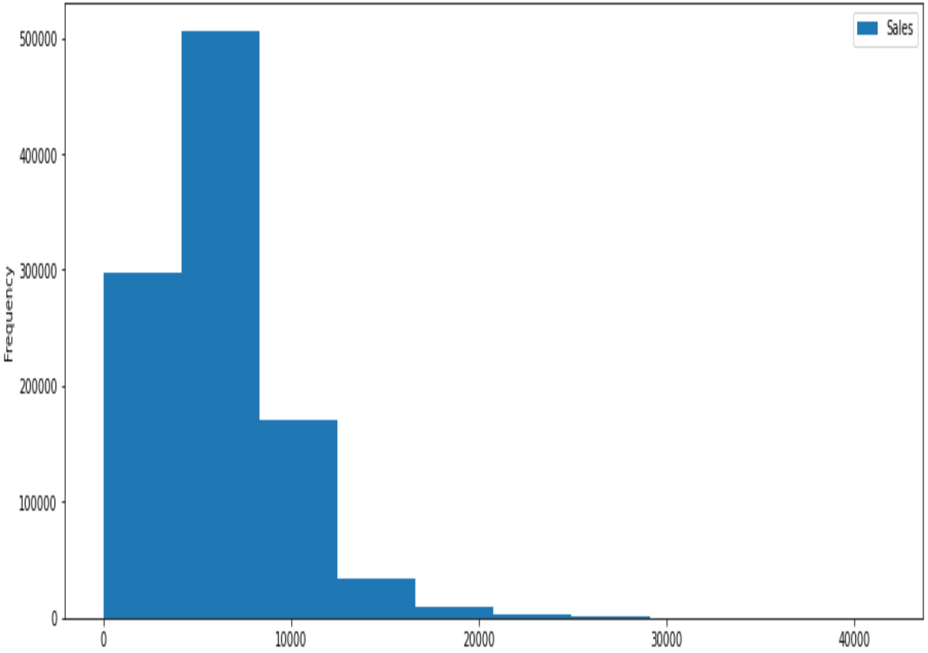
EDA

Competition Distance vs Sales:

As the Competition distance increases, the sales tend to increase due to monopolization in that certain area.

The trend starts to go down as the distance increases from 8000m.

This might be indicating the stores located at remote location with less dense population, where the competition is far but customers are low anyways.

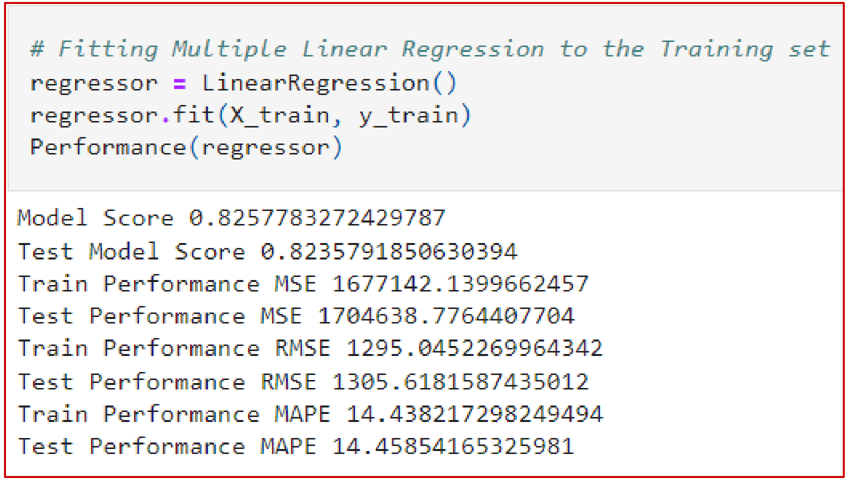


Feature Engineering and Selection

- Now that we understand the relation and effects that some features have on sales and how the data is represented, we proceeded with elimination of features with high null values and less relevance.
- State holiday feature has nominal values which we converted to binary as, we just wanted to know if there was a Holiday.
- We had a lot of records with zero sales of new stores and refurbishment stores, which ended up getting dropped.
- Got dummies for some categorical values like storeype, assortment and year.

Utility Functions

After Feature selection, we proceeded to split them into dependent and independent variables, which was then followed by train and test splitting.
Before we fed the data to the models, we needed some utility functions to reduce clutter in the further phase. Creation of a utility function called MAPE( ) was done for calculating the Mean Absolute Percentage Error.
Another utility function calculated the performance of models both training and testing followed by error scores. This function was Performance(model_name).



```python
# Fitting Multiple Linear Regression to the Training set
regressor = LinearRegression()
regressor.fit(X_train, y_train)
Performance(regressor)
```

```
Model Score 0.8257783272429787
Test Model Score 0.8235791850630394
Train Performance MSE 1677142.1399662457
Test Performance MSE 1704638.7764407704
Train Performance RMSE 1295.0452269964342
Test Performance RMSE 1305.6181587435012
Train Performance MAPE 14.438217298249494
Test Performance MAPE 14.45854165325981
```

Output of the Performance() function.

Model Fittings and Feature Importance

In order to predict the continuous target variable sales, regression models can be of help. In this project, four different regression models were used with hyperparameter tuning to get the best predictive model.

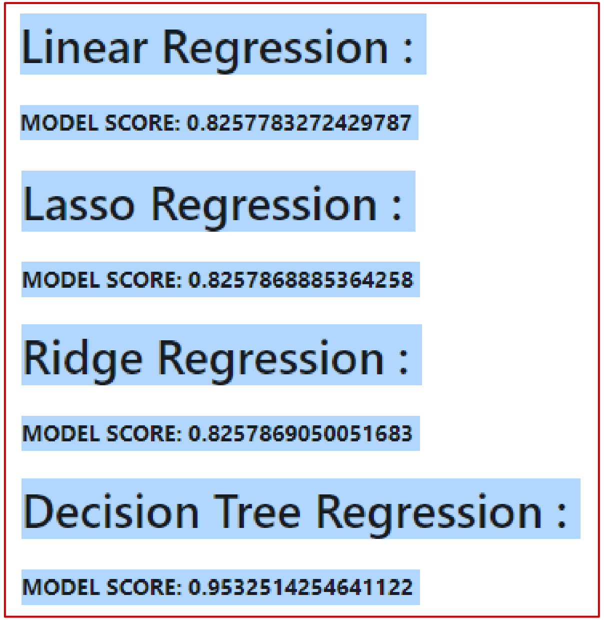
We used –

- Linear Regression
- Lasso Regression
- Ridge Regression
- Decision Tree Regressor

Model Fittings and Feature Importance

These were the scores that every model generated. The scores are training model scores. Linear Regression model performed really well but not as well as Lasso and Ridge regression. The difference is fairly low but even the tiniest improvement counts.

The best model turned out to be Decision Tree Regressor.

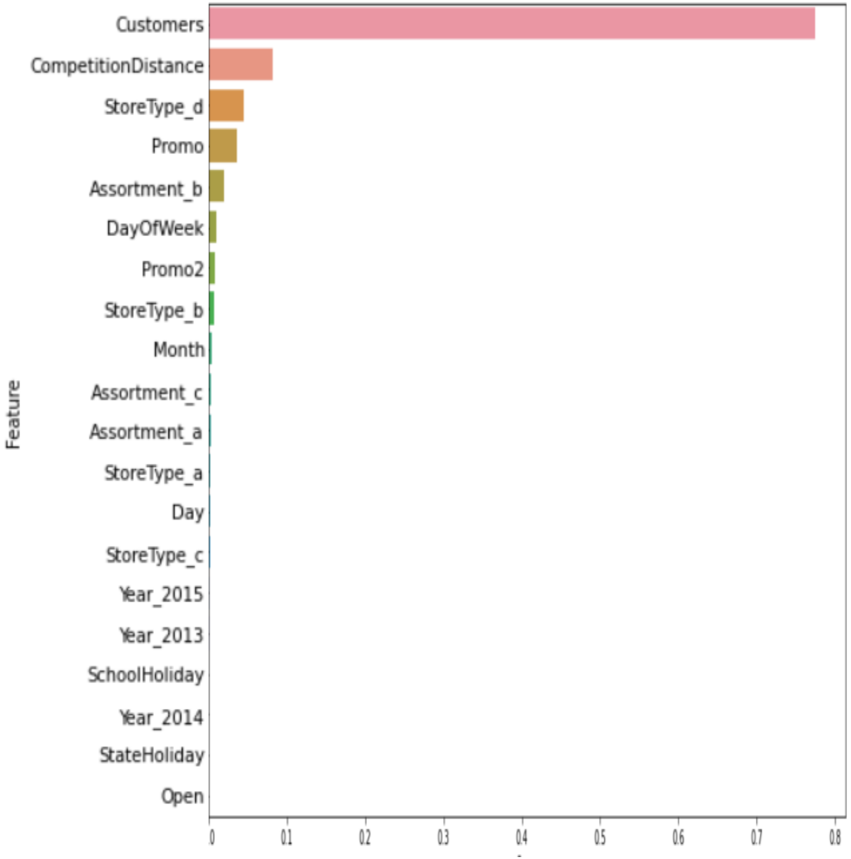


```
Linear Regression :

MODEL SCORE: 0.8257783272429787

Lasso Regression :

MODEL SCORE: 0.8257868885364258

Ridge Regression :

MODEL SCORE: 0.8257869050051683

Decision Tree Regression :

MODEL SCORE: 0.9532514254641122
```

Model Fittings and Feature Importance

In Decision Tree Model, these were the features which were given the most importance in terms of driving sales. We can ignore the feature ‘customers’ as it is obvious.

Features like Competition distance, Store Type, Assortment and Promo contribute towards sales the most.



Challenges Faced

There were 3 major challenges which effected the overall analysis the most.

- Missing Values of Competition details. (almost 48% missing data)
- Missing Values of Promo2 details. (almost 31% missing data)
- No proper description of Store type and Assortment.

All these missing and improper data description lead to assumption based analysis. The meaning of Basic, Extra and Extended cannot be comprehended. All we can report is, the sales to customer ratio and their consistency.

Conclusion

- Customer and Sales show a linear growth, indicating sustained customers.
- Promotional offers help in sales but not significantly.
- Competition distance, assortment, store type and promo are the features which drive the sales.
- Decision Tree Regressor model turned out to be the best among tested predictive models.
- The testing model score for DTR was 0.93 which is very high.