Capstone Project - 3

Supervised ML - Algorithms

on

Retail Sales Prediction

By – Rishanshu Yadav

❑ CONTENT -

1. Problem Statement
2. Data Summary
3. Data Preprocessing
4. Exploratory Data Analysis
5. Feature Engineering
6. Model Implementation
7. Conclusion

❑PROBLEM STATEMENT -

Rossmann operators over 3,000 drug stores in 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 set . Note that some stores in the dataset were temporarily closed for refurbishment

❑DATA DESCRIPTION -

Rossmann Stores Data.csv - historical data including Sales

Store.csv – supplemental information about the stores

{ After merging both the dataset we have 1017209 number of records and 16 number of fields and our dataset period is from 1st jan 2013 to 31st july 2015.}

❑DATA FIELDS -

- Most of the fields are self-explonatory . The following are description for those that are't .

1. **Customer** : - The number of customers on a given day in a store.
2. **Date** :- Showing dates for observations.
3. **State Holiday** :- Indicating a state holiday. Normally all stores , with few
expectations , are closed on state holidays . Note that all
schools
are closed on public holidays and weekends .

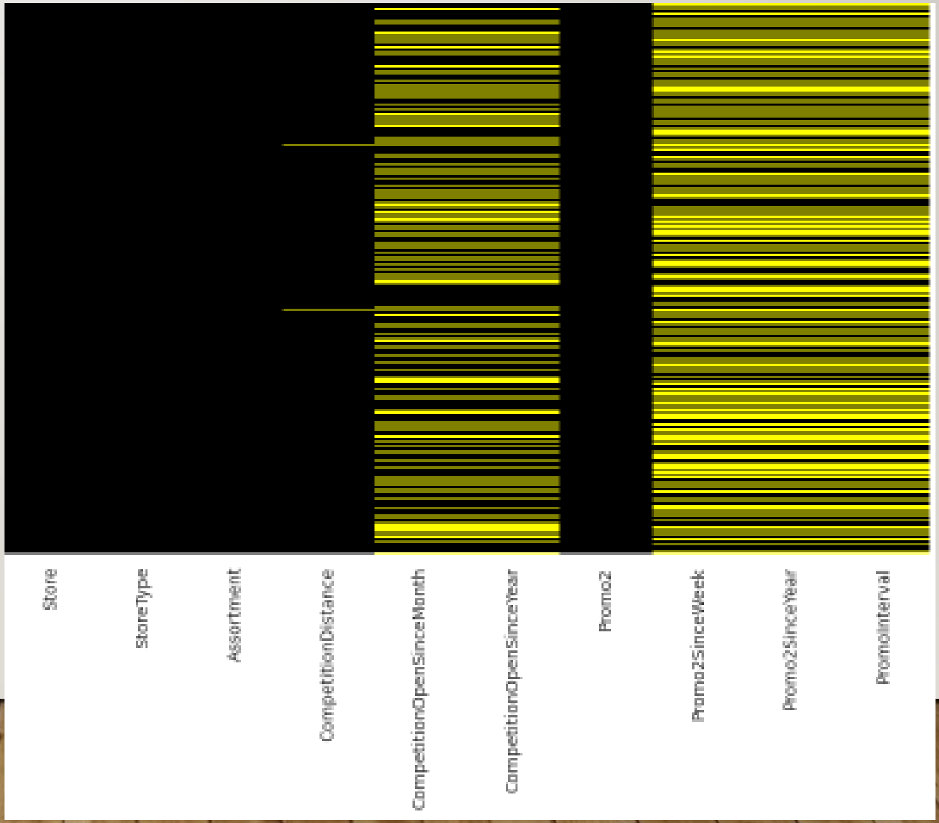
a = public holiday , b = easter holiday , c = Christmas , O = None

4. **Store Type** : Differentiate between 4 different store models (a,b,c,d).
5. **Assortment** : Describes an assortment level i.e a : basic, b : extra and c : extended.
6. **Competition Distance** : Distance in meters to the nearest competition store.
7. **Promo** :- Indicates whether a store is running a promo on that day
8. **Open** :- the number for whether the store was open; 0 = closed , 1 = open.
9. **School Holiday :**– indicates if the (Store , Date) was affected by the closure of public schools
10. **CompetitionOpenSince[Month,year]** – gives the approximate year and month of the time the nearest competitor was opened.

❑ DATA PREPROCESSING -

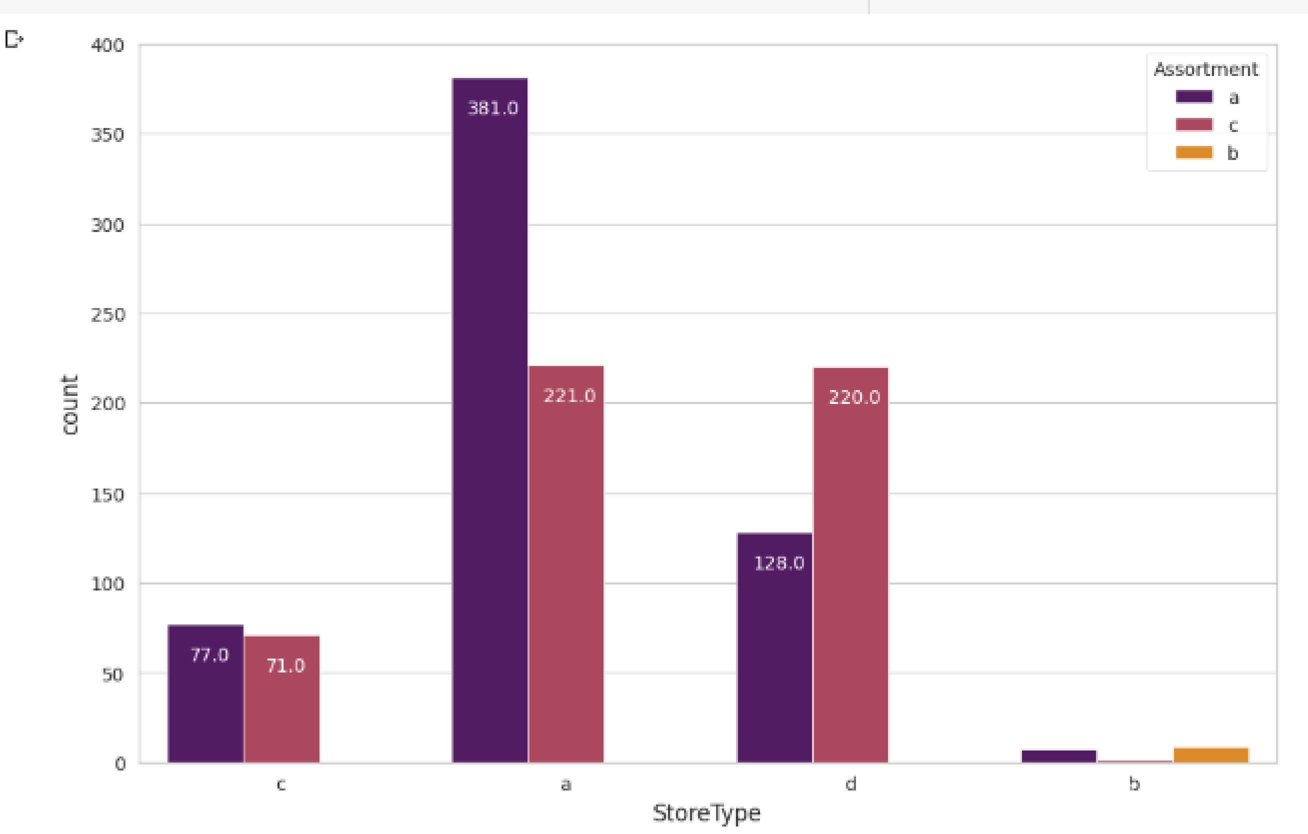
- ➢ Data wrangling and processing required cleaning of data and preparing it for further analysis . Our process involved the follwing parts :

- We have merge both the available dataset.
- Columns having >30% null values are dropped.
- Null values in ‘Competition Distance’ are imputed with median of feature.
- Removing those stores observations that are temporarily closed (~ 17.3K) & stores generating zero sales



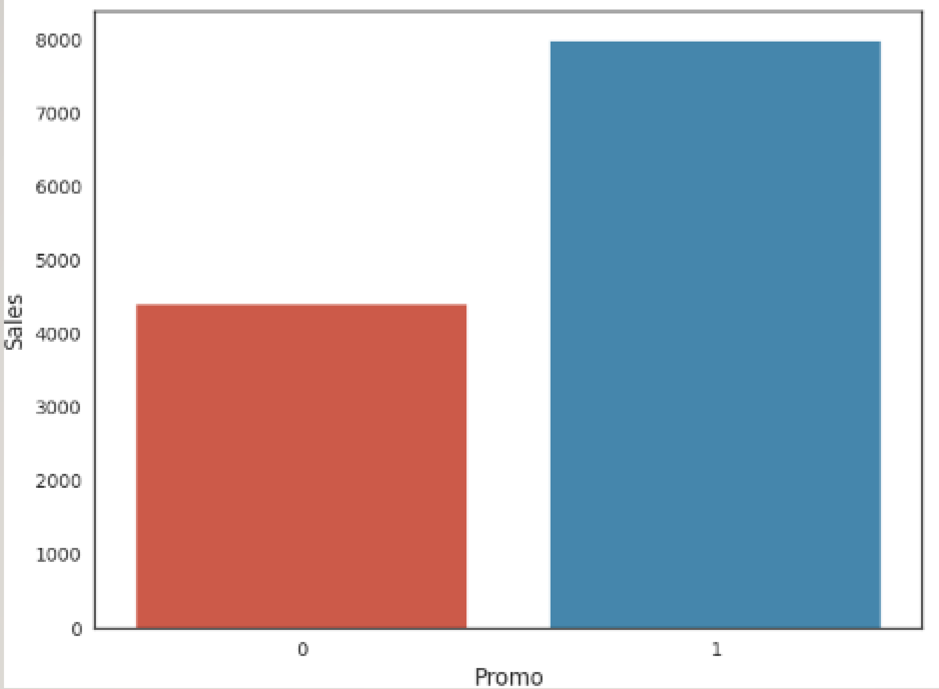
❑ EXPLORATORY DATA ANALYSIS (EDA)

❑ Analysis of Store Types with their respective assortment -



❖ Store type 'a' have the maximum number of sales and store counts followed by 'd' while store 'c' and store 'b' have the least number of sales and store counts

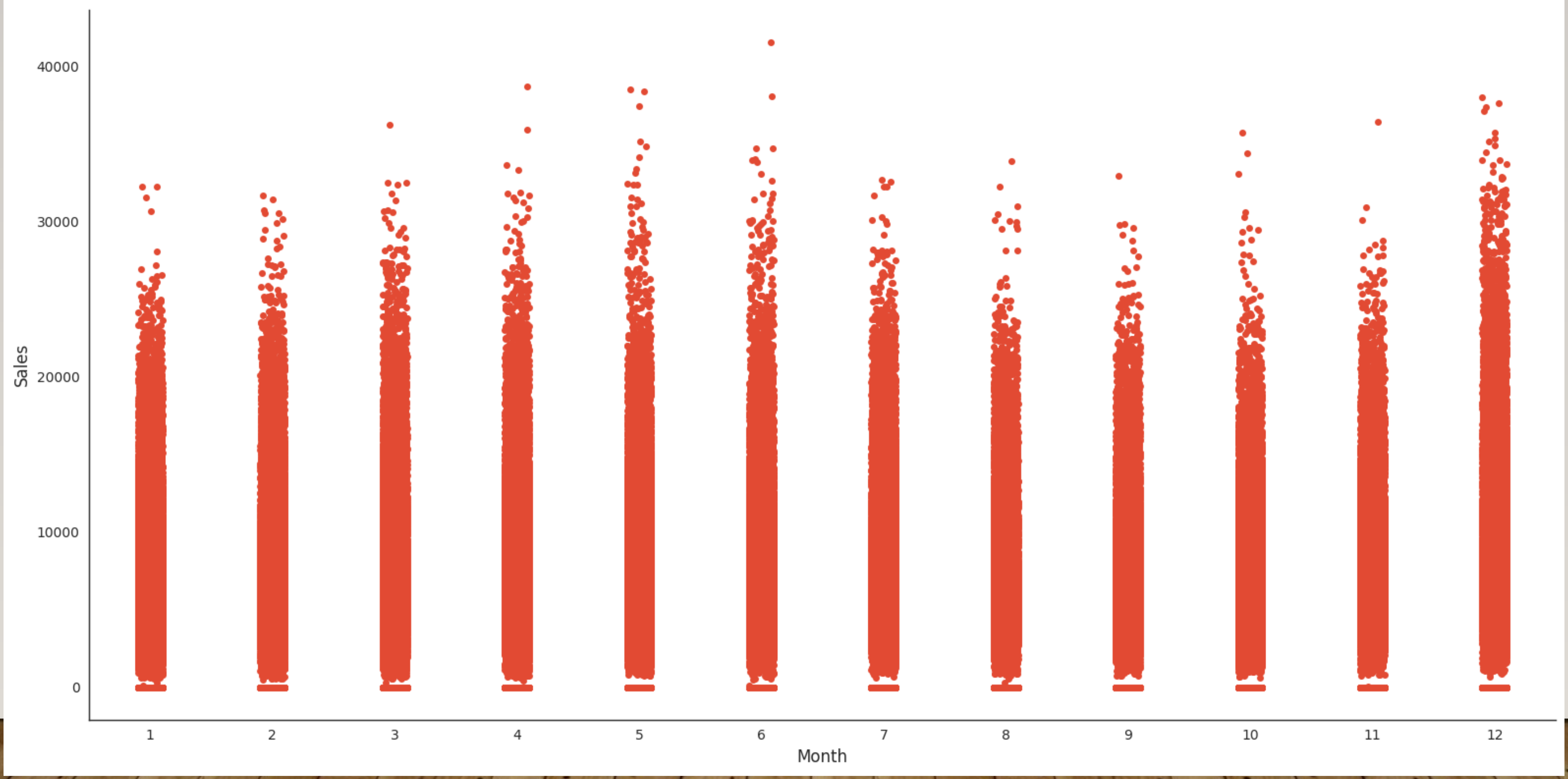
❑ IMPACT OF PROMO ON SALES AND CUSTOMER



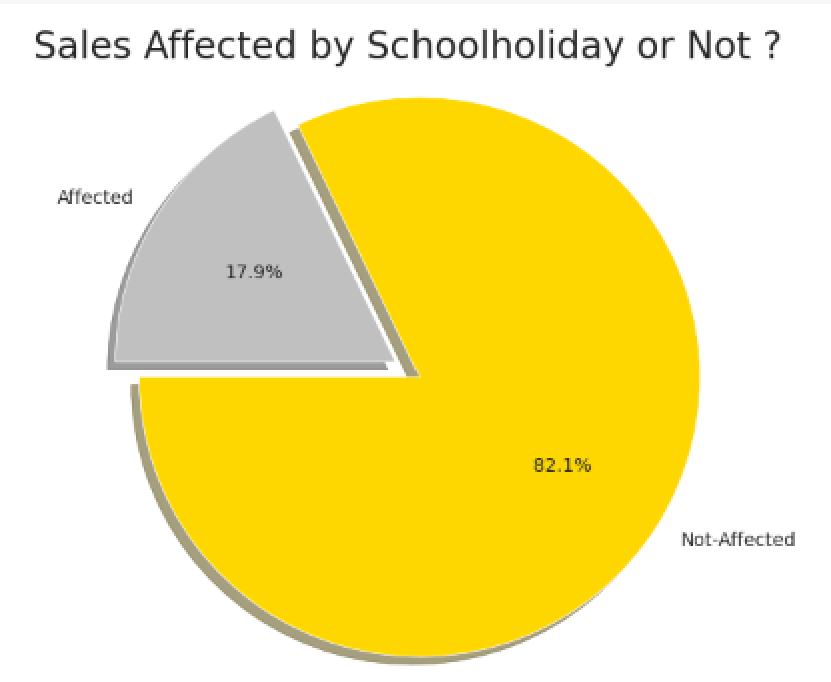
❖There is a linear relationship between customers and sales and it also noticeable that Whenever the promo was open , we can see that the sales almost doubled to the , Whenever the promo was closed , which means promo had good impact on the Business.

❑ SALES ON EACH MONTH -

- As we can see Sales is at peak during december and april , may and june while sales is at lowest during january ,Anugust and september .

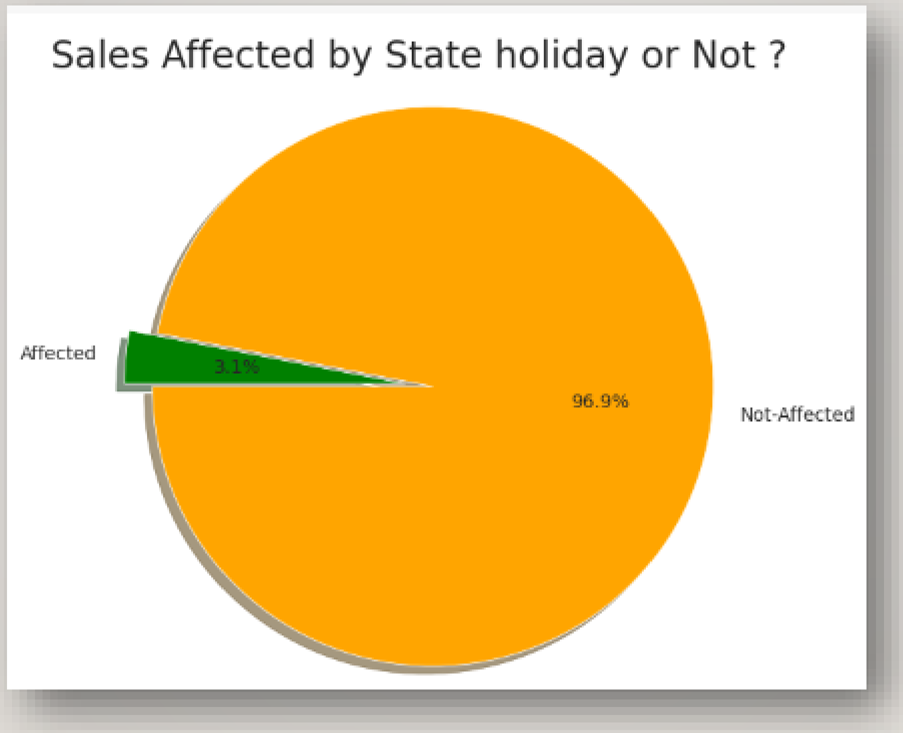


❑ IMPACT OF SCHOOL HOLIDAYS ON SALES -



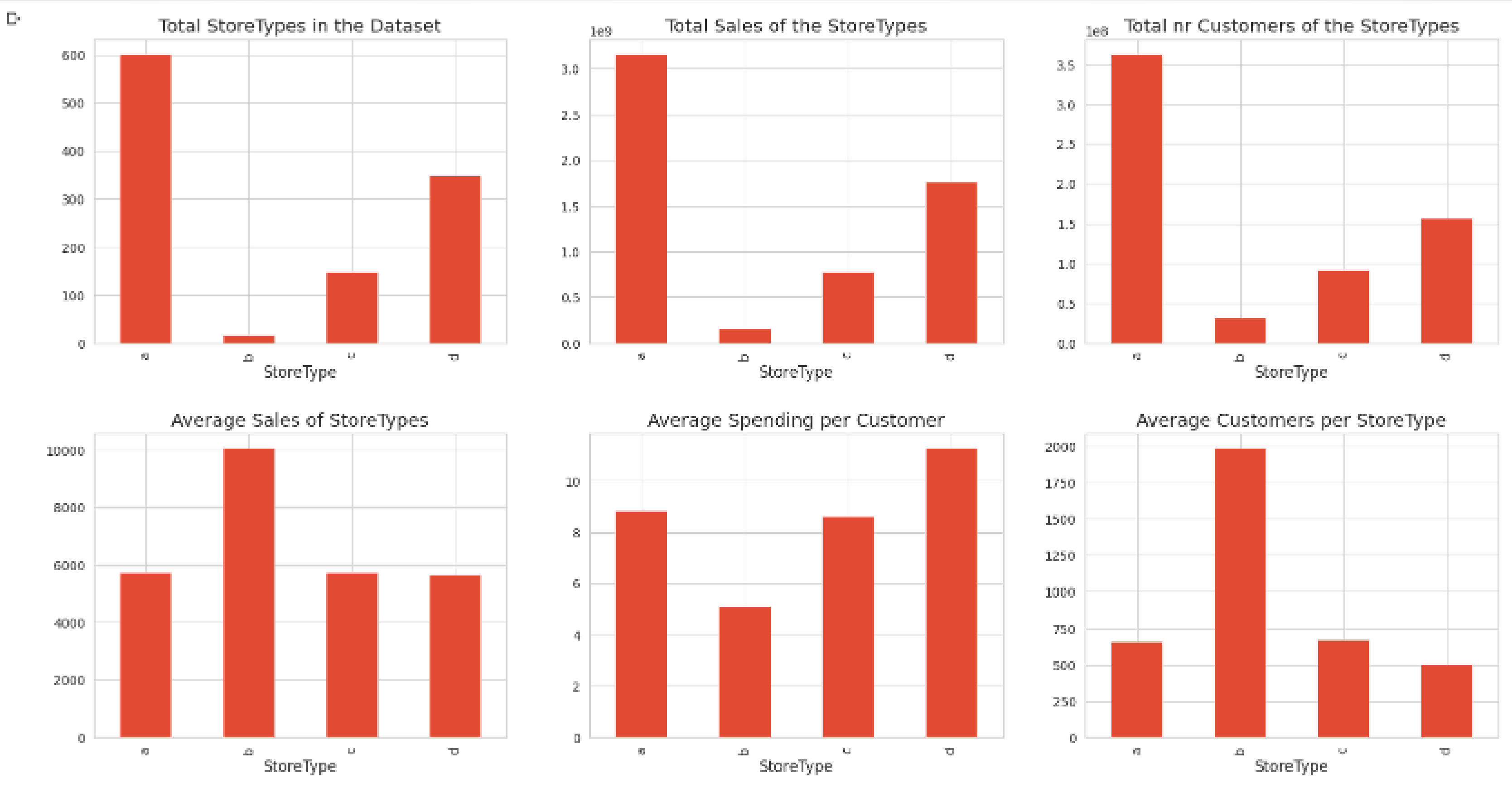
❖ As we see , 17.9% of the total sales gets affected by the school holidays and 82.15 were not .Which also mean that around 17% of the sales are oriented from the school students .

❑ IMPACT OF STATE HOLIDAY ON SALES -



- ❖ As we see, 3.1% of the total sale is affected by state holiday and 96.9% sale is not affected by state holiday.

- Store Types and average sales/customer/spending relation



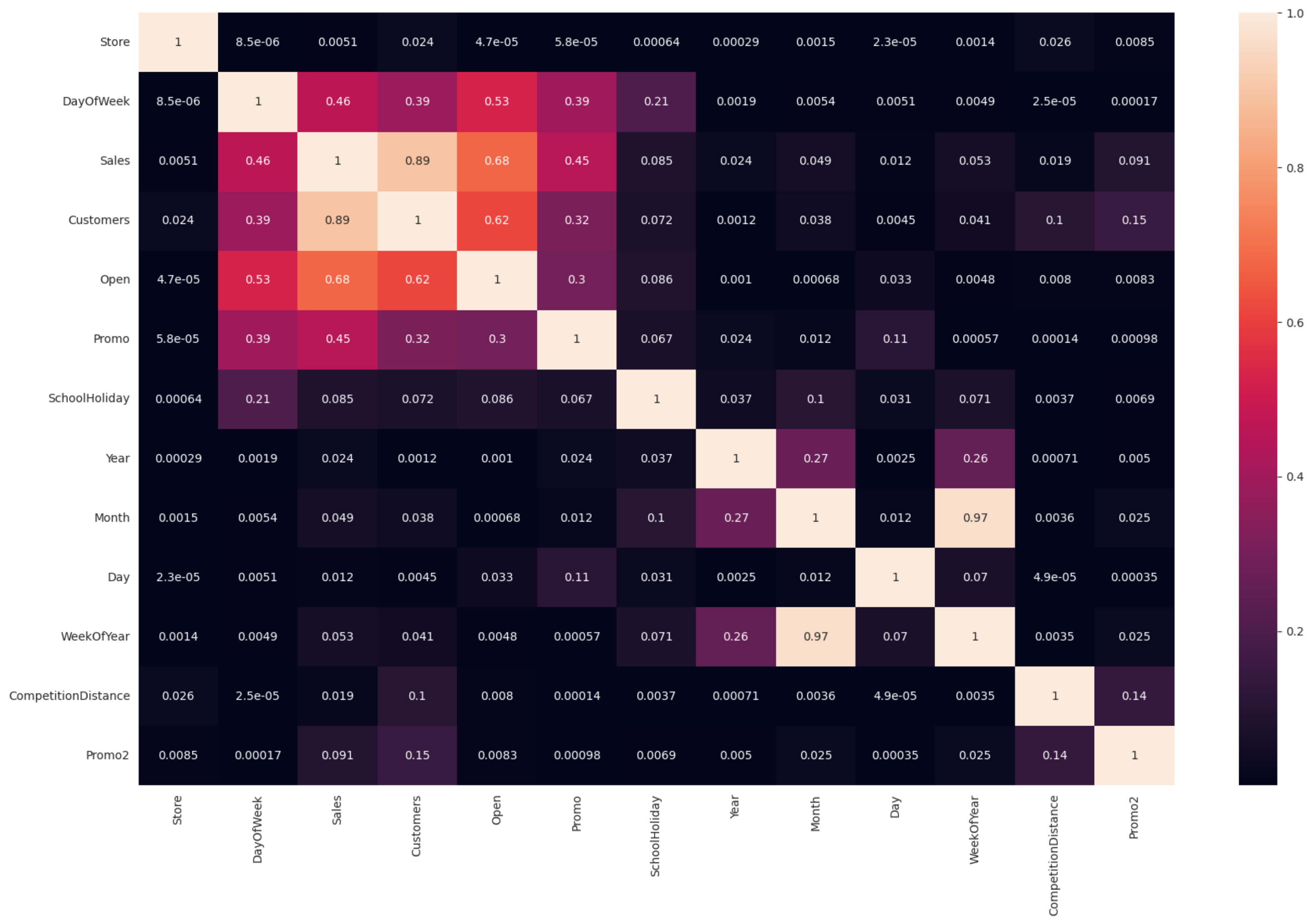
❑ FEATURE ENGINEERING -

- Before procceding to ML Model we did some features engineering to simply and speed up data transformation while also enhancing model accuracy.

- **Feature Selection** – After analysis each column we find out that ‘Store’ column have all unique values which won’t help me in any prediction so we dropped.

- **Scaling Numerical Feature**: We used MinMaxScaler to scale numerical variables **[‘Customer’,‘CompetitionDistance’ , ’Year’,’Month’, ‘DaysOfWeek’,’Days’,‘CompetitionOpen’,’PromoOpen’]** within a given range of 0 to 1 .

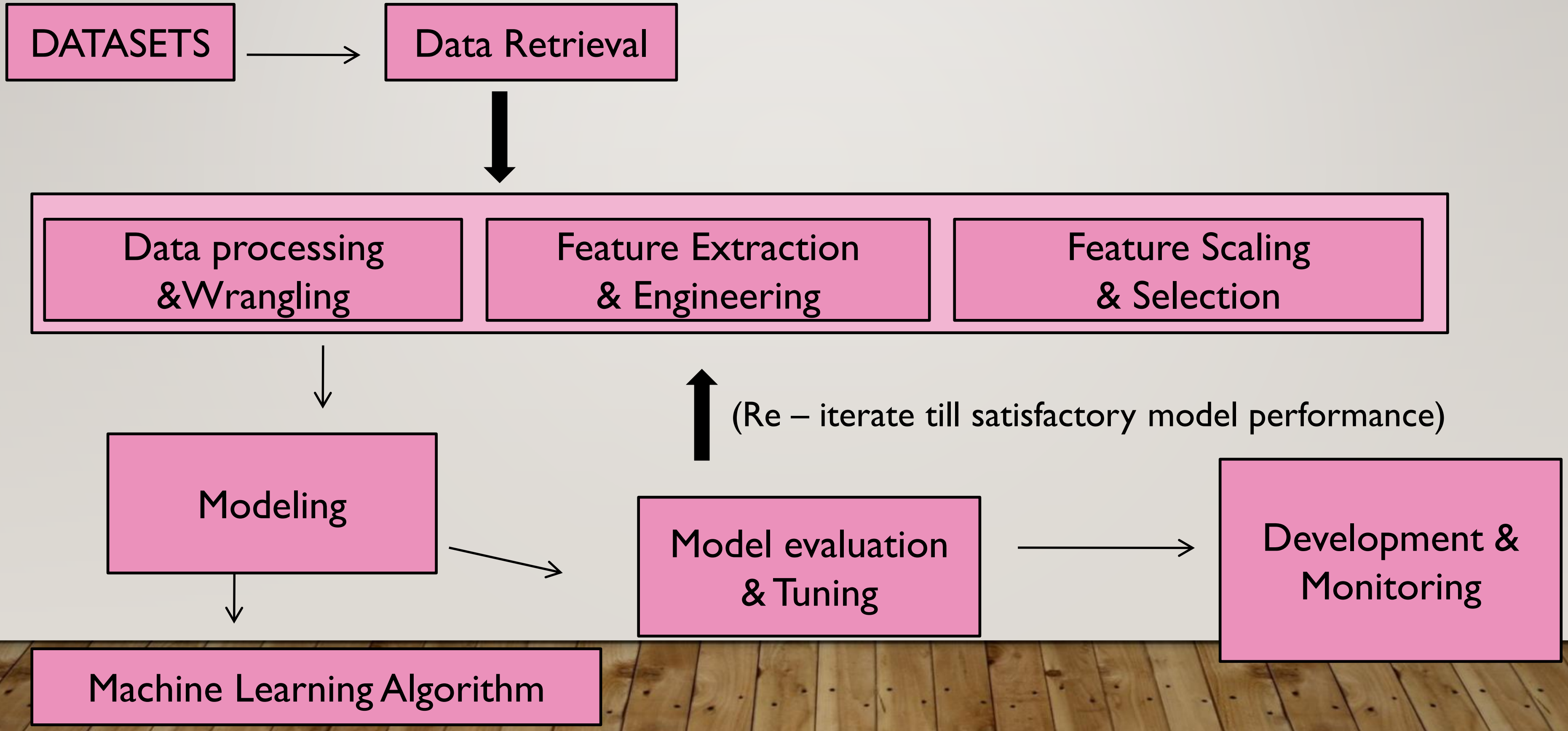
➢**Multicollinearity: -** We didn’t find any correlation between independent Variables but we found some correlation with our dependent features which is a good sign for our model.



	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday	Year	Month	Day	WeekOfYear	CompetitionDistance	Promo2
Store	1	8.5e-06	0.0051	0.024	4.7e-05	5.8e-05	0.00064	0.00029	0.0015	2.3e-05	0.0014	0.026	0.0085
DayOfWeek	8.5e-06	1	0.46	0.39	0.53	0.39	0.21	0.0019	0.0054	0.0051	0.0049	2.5e-05	0.00017
Sales	0.0051	0.46	1	0.89	0.68	0.45	0.085	0.024	0.049	0.012	0.053	0.019	0.091
Customers	0.024	0.39	0.89	1	0.62	0.32	0.072	0.0012	0.038	0.0045	0.041	0.1	0.15
Open	4.7e-05	0.53	0.68	0.62	1	0.3	0.086	0.001	0.00068	0.033	0.0048	0.008	0.0083
Promo	5.8e-05	0.39	0.45	0.32	0.3	1	0.067	0.024	0.012	0.11	0.00057	0.00014	0.00098
SchoolHoliday	0.00064	0.21	0.085	0.072	0.086	0.067	1	0.037	0.1	0.031	0.071	0.0037	0.0069
Year	0.00029	0.0019	0.024	0.0012	0.001	0.024	0.037	1	0.27	0.0025	0.26	0.00071	0.005
Month	0.0015	0.0054	0.049	0.038	0.00068	0.012	0.1	0.27	1	0.012	0.97	0.0036	0.025
Day	2.3e-05	0.0051	0.012	0.0045	0.033	0.11	0.031	0.0025	0.012	1	0.07	4.9e-05	0.00035
WeekOfYear	0.0014	0.0049	0.053	0.041	0.0048	0.00057	0.071	0.26	0.97	0.07	1	0.0035	0.025
CompetitionDistance	0.026	2.5e-05	0.019	0.1	0.008	0.00014	0.0037	0.00071	0.0036	4.9e-05	0.0035	1	0.14
Promo2	0.0085	0.00017	0.091	0.15	0.0083	0.00098	0.0069	0.005	0.025	0.00035	0.025	0.14	1

❑ML MODEL -

After performing all these steps our dataset is ready for ML Modeling . Now we will train a Model over a set of data , providing it an algorithm that it can use to reason over and learn From those data and then making predictions on those data which hasn't been seen.



DATASETS

Data Retrieval

Data processing &Wrangling

Feature Extraction & Engineering

Feature Scaling & Selection

(Re – iterate till satisfactory model performance)

Modeling

Model evaluation & Tuning

Development & Monitoring

Machine Learning Algorithm

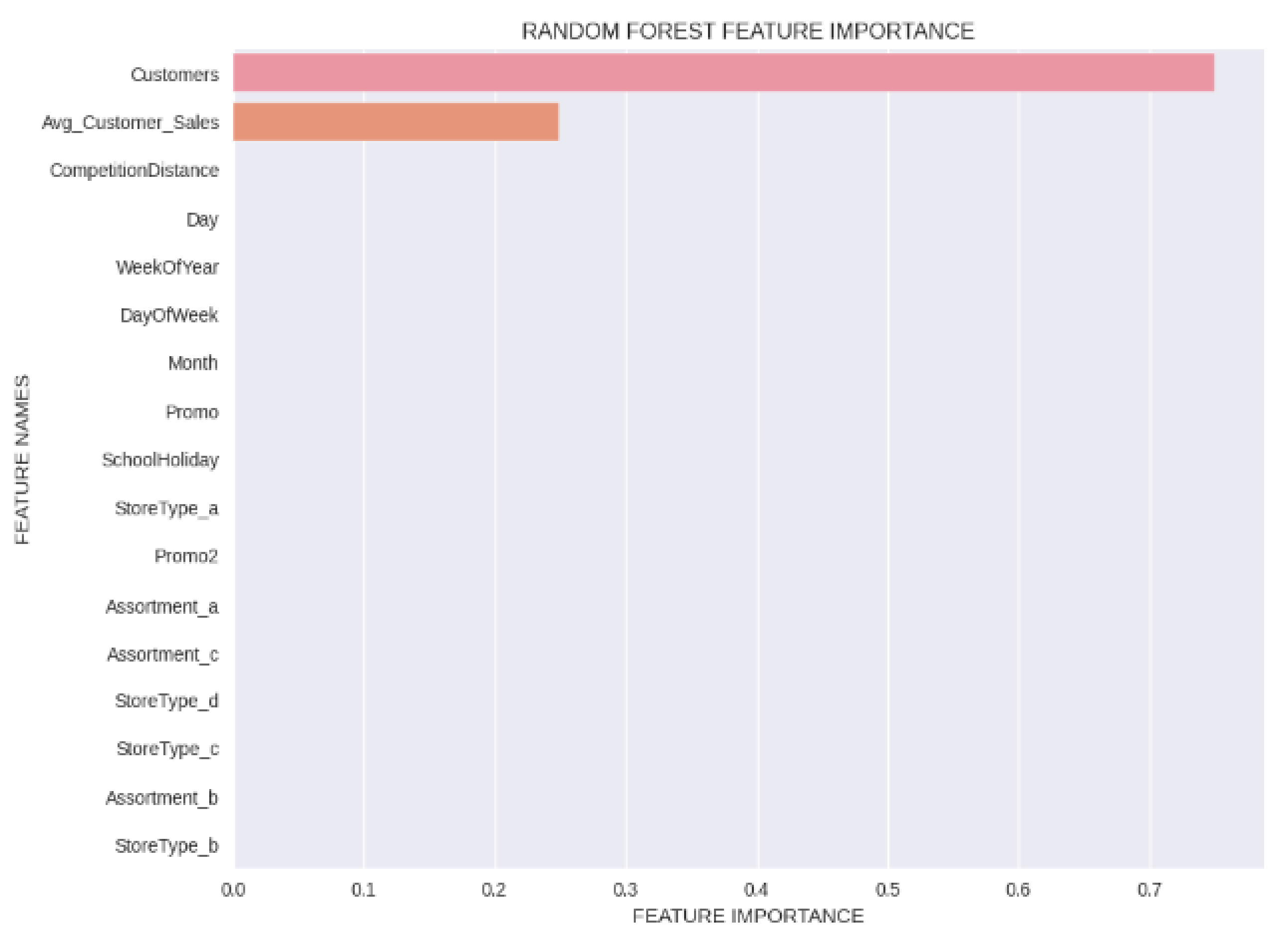
❑ ML Model Performance -

➢ After performing various regression techniques on our predictive model and we found Out that Random Forest Regression has performed better than any other regression Model .

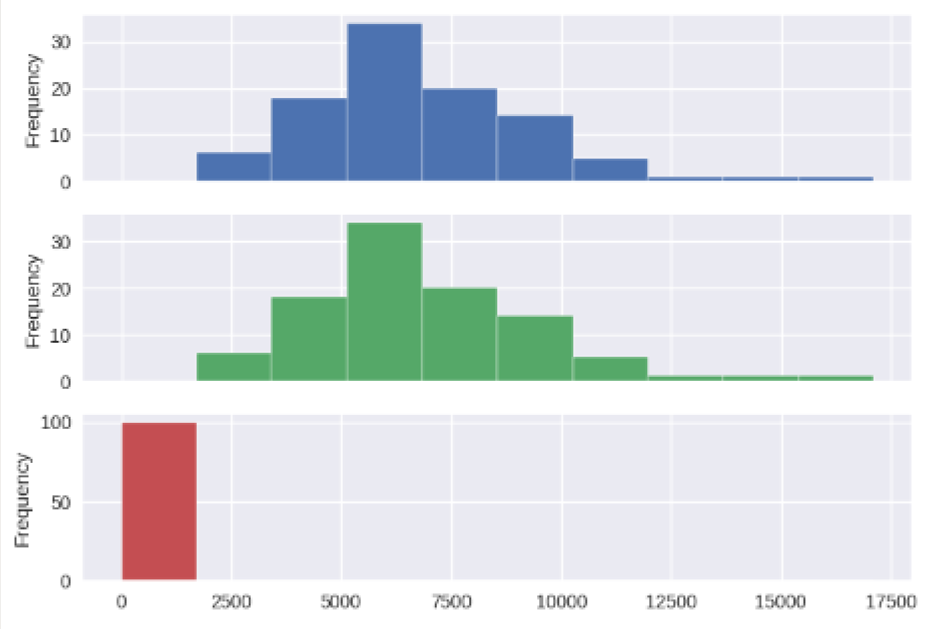
MODEL	TRAINING SCORE	TESTING SCORE
Linear Regression	0.780750	0.782392
Lasso Regression	0.780731	0.780769
Decision Tree Regression	0.99996	0.915942
Decision Tree Regression (with hyper parameters)	0.963506	0.935417
K-Nearest Neighbours Regression	0.73722	0.71665
Random Forest Regression	0.993783	0.956520

❑FEATURE IMPORTANCE -

❖ After selecting our regression model we can see the importance of each features in our model prediction



❑Insights from Random Forest Regressor -



- 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 ~ $ 380 and MAPE of 5.65%

❑ CONCLUSION -

- Store model ‘b’ have least number of stores in Rossmann yet it performed well and made more sales than other store model so it is advisable to increase the number of ‘b’ store model.
- Assortment level ‘Basic’ have the maximum number of stores in Rossmann yet it performed very badly but at the same time ‘Extra’ and ‘extented’ assortment level with less number of store had performed extra ordinary so it would be adviseable to increase these asssortment level.
- Linear relationship have been found among customers , sales and promo . And it has been seen that most of the customers came for shopping during the promo days as the cost was lower on those days .

  So, promo should be initiated to more stores to increase the sales.
- Sales has been low on the initial days of the month as compared to the end days , it can be assumed that people used to shop for the next month at the end of the previous month . These products can be mainly be of basics of a person’s daily life .

THANK YOU