Supervised ML – Regression

Capstone Project

Retail Sales Prediction

Team Members–

1. Tushar Khairnar
2. Shreyash Sarode
3. Pradnya Pagar
4. Taha Nakedar

Problem Statements

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.

Points of Discussion

1. Problem Statements
2. Understanding Dataset
3. Data Pre-Processing
4. Exploratory Data Analysis
6. Feature Engineering
5. ML Model
6. Feature Importance
7. Challenges faced
8. Conclusions

Understanding The Dataset

Data Description

Rossmann Stores Data.csv - historical data including Sales

store.csv - supplemental information about the stores

{After merging both the datasets we have 1017209 number of records and 18 number of fields and our dataset period is from 1st Jan-2013 to 31st July-2015.}

Data Fields

Most of the fields are self-explanatory. The following are descriptions for those that aren't.

- **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 we are predicting)

Customers - the number of customers on a given day.

Open - an indicator for whether the store was open: 0 = closed, 1 = open.

StateHoliday - 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

SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools

StoreType - differentiates between 4 different store models: a, b, c, d

Assortment - describes an assortment level: a = basic, b = extra, c = extended

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:

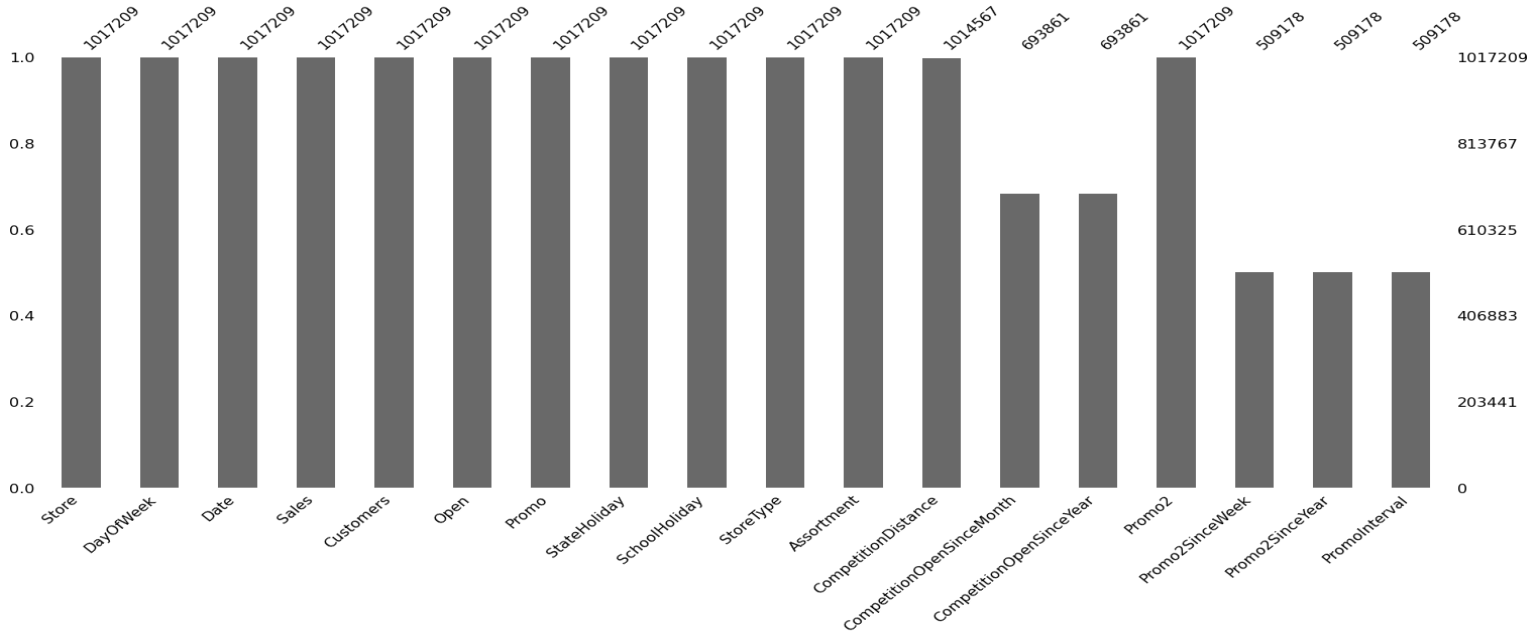
0 = store is not participating, 1 = store is participating

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

Data Pre- Processing

As We have a Dataset of Rossmann Stores which contain 1017209 rows and 18 columns. Some columns have missing values.



Data Pre- Processing

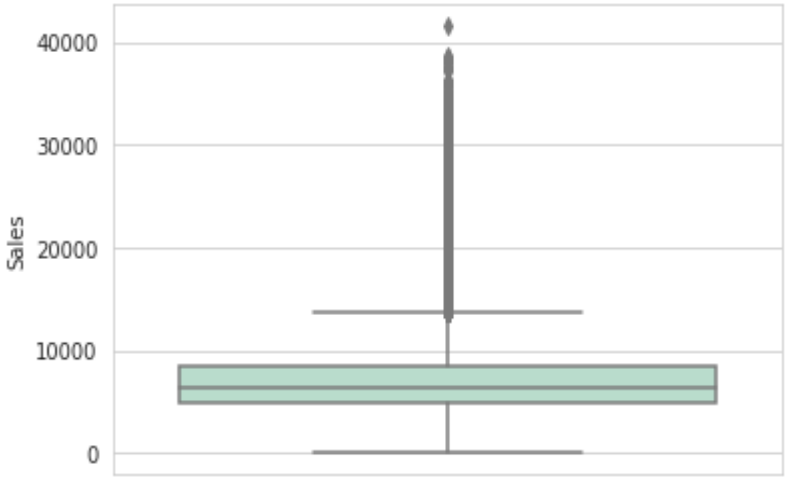
Data wrangling and processing requires cleaning of data and preparing it for further analysis. Our cleaning process involved the following parts:

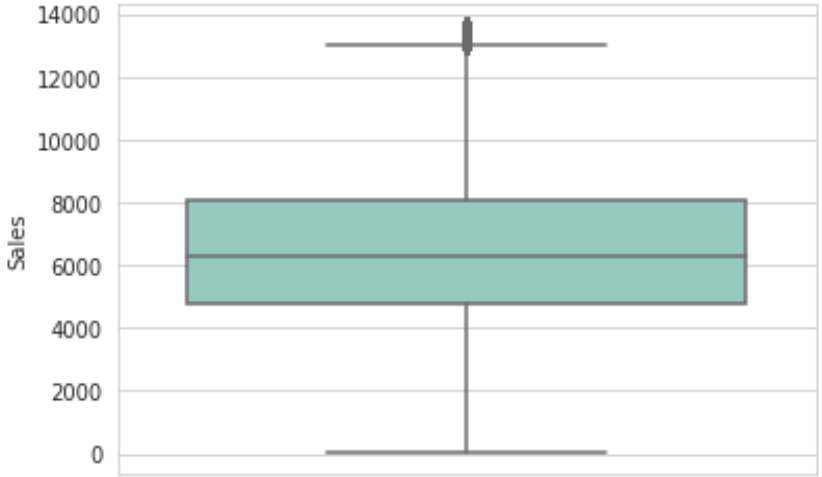
- **Merge Both Dataset:** We have merge both the available dataset

- **Null Value Treatment:**
  - After checking distribution of **'CompetitionDistance'** column we find out that data is left skewed so we use median to replace null values.
  - And we replace the null values present in column **'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear'** with mode.
  - We have dropped columns where null values present in **'Promo2SinceWeek', 'Promo2SinceYear', 'PromoInterval'** because columns having more number of null values and these columns are not much impactful.

Handling Outliers in Target Variable

- Started with our target variable as it is the most important variable.
- This Data set have some genuine values which seems as outliers. so we had worked on only those values which are very important to remove and removal of those will does not affect our data set.





Exploratory Data Analysis

Basically we have two important categorical columns which need explanation in our dataset so lets start our visualization with those data.

Assortment

- a = Basic
- b = Extra
- c = Extended

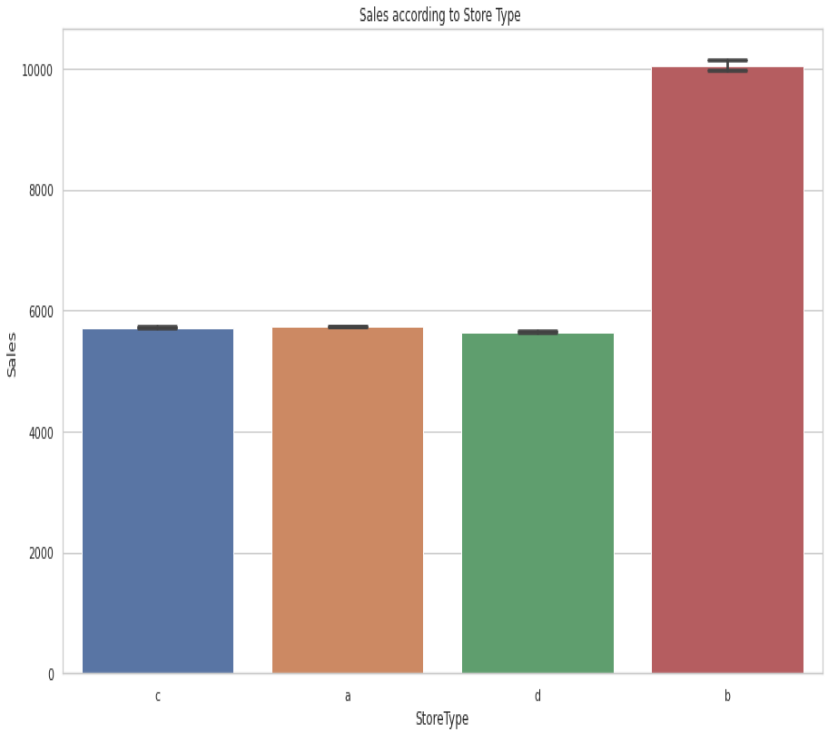
State Holidays

- a = Public Holiday
- b = Easter Holiday
- c = Christmas Holiday
- d = None

EDA

Store Models

- ✓ This plot shows relation between Store Type and Sales.
- ✓ In the above chart The sales in the store type "b" have higher sales as compare to other store category. the rest of store having almost same sales.
- ✓ If you want more sales then go with "b" Store Type.

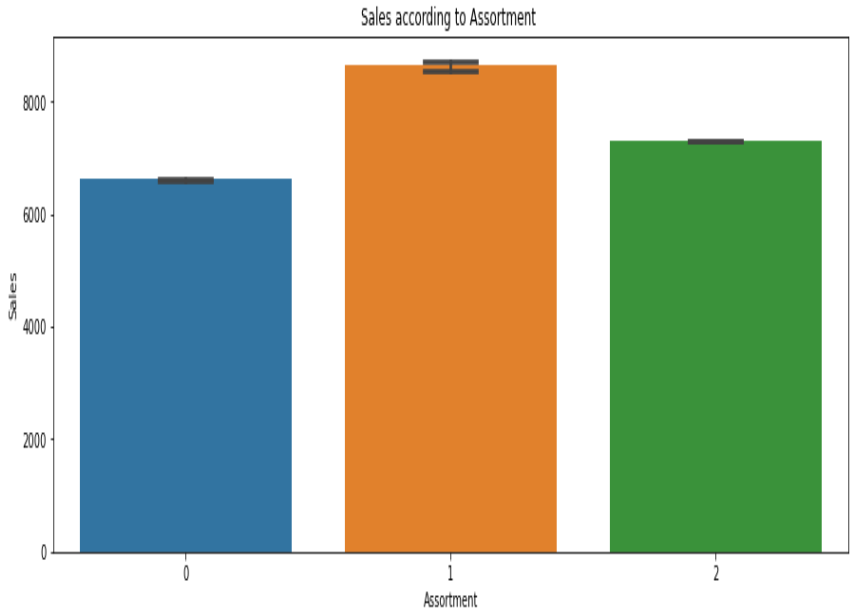


EDA

Assortment Levels

0 = Basic, 1 = Extra, 2 = Extended

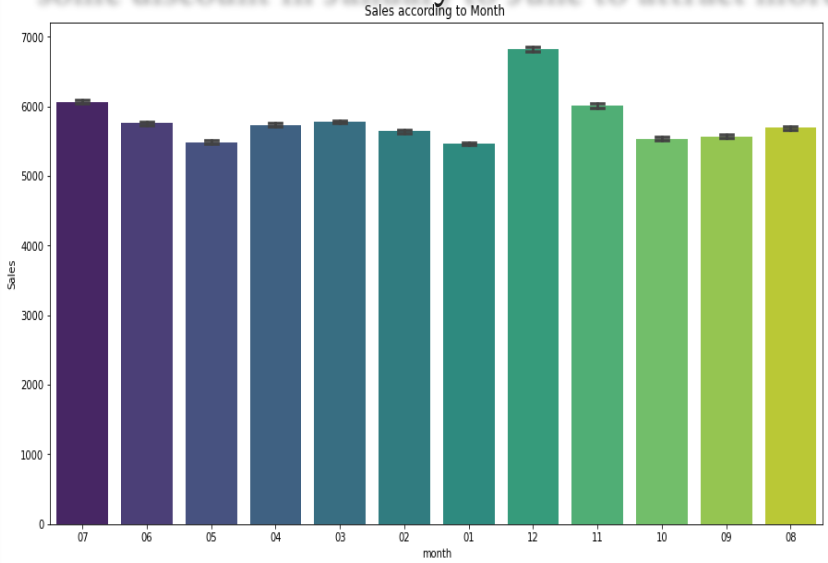
- The sales in which product assortment type 1 have higher sales as compare to other assortment category. the rest of assortment having almost same sales.

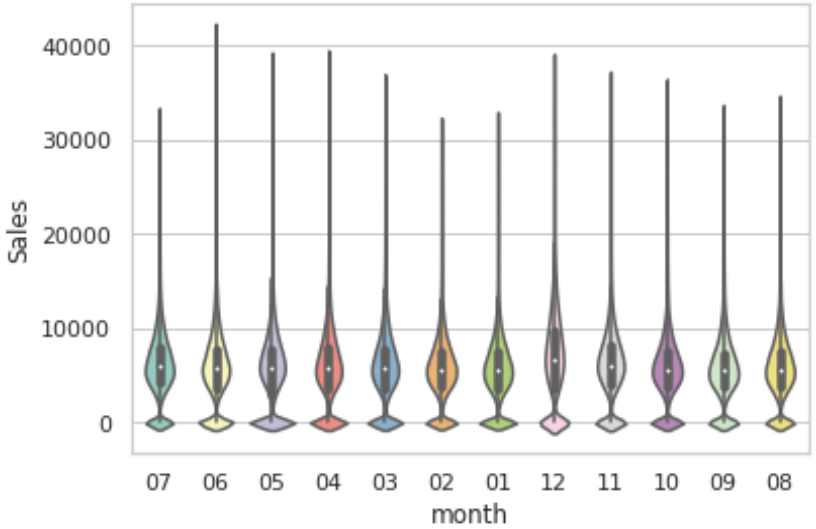


EDA

Sales in each Month

According to the plot we can see that the sale is increases in the December and Decreases in January and we can conclude that stores need more supply in between July to December stores should offer some discount in January to June to attract more customers.

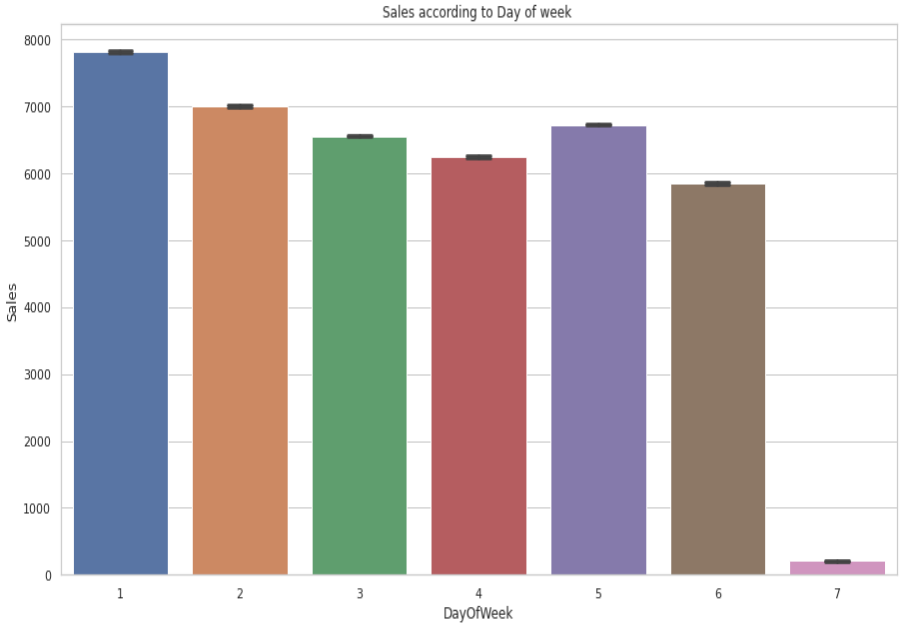




EDA

Days of week with Sales

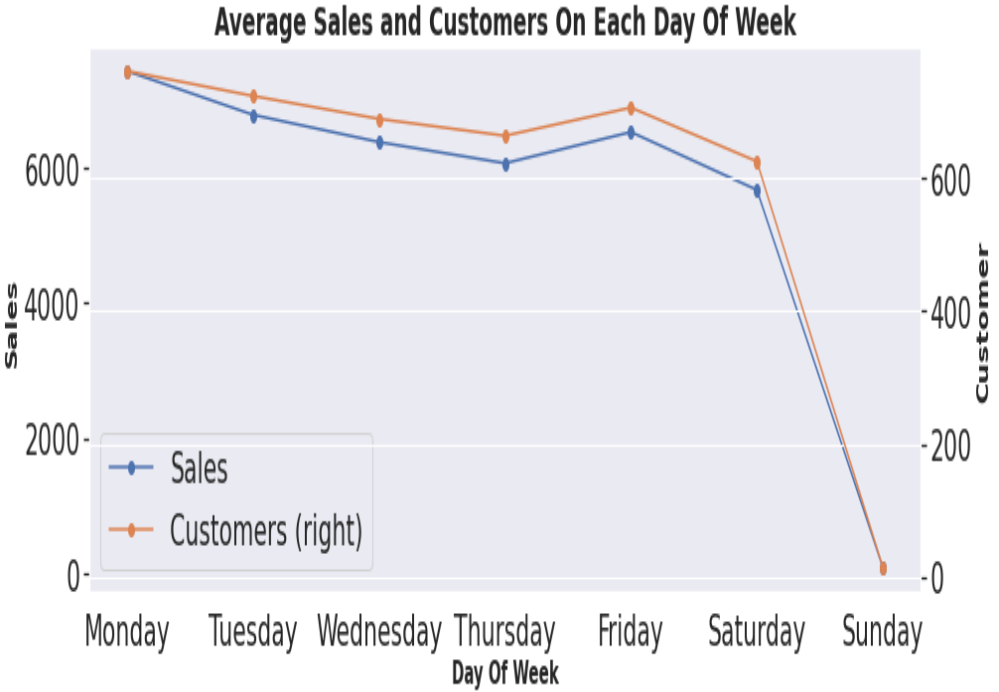
- We are plotting Bar graph to see the Sales according to Days in week.
- We can conclude that sales are high on Monday and low on Sunday.
- It may be because of holiday Stores have to try increasing sale on sunday or they can close the store for half day on sunday to minimize expenditure.



EDA

Average Sales And Customers On Each Day Of Week

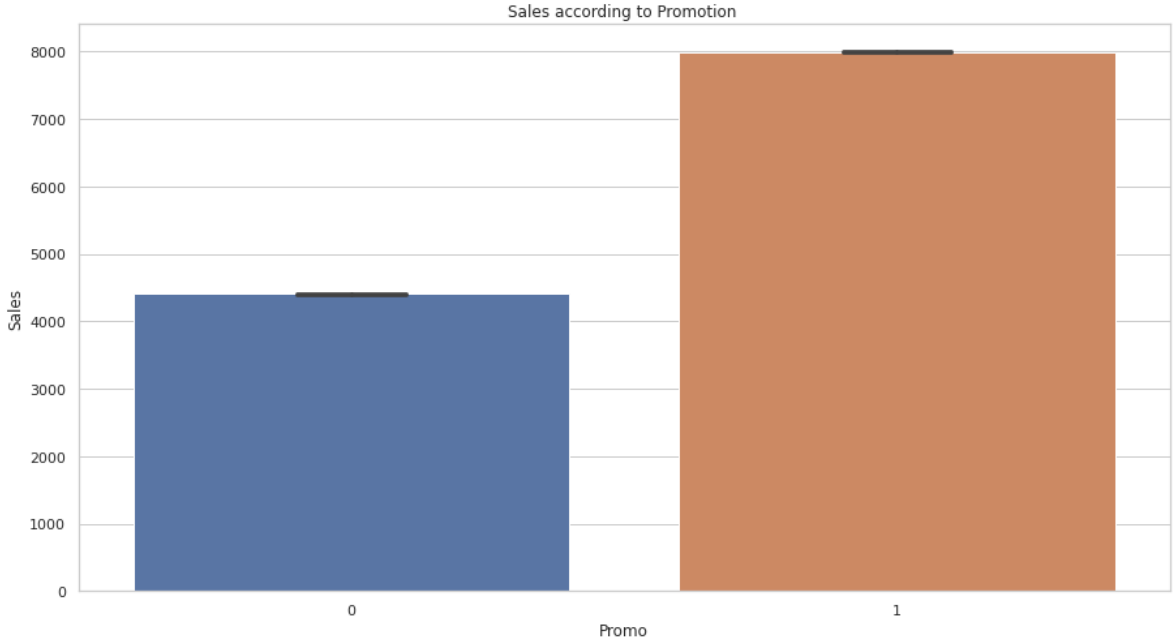
Sales and customers are at maximum on Mondays while sales and customers are nearly zero on Sundays because it seems like store use to remain closed on Sundays.



EDA

Prices In Neighbourhood Groups

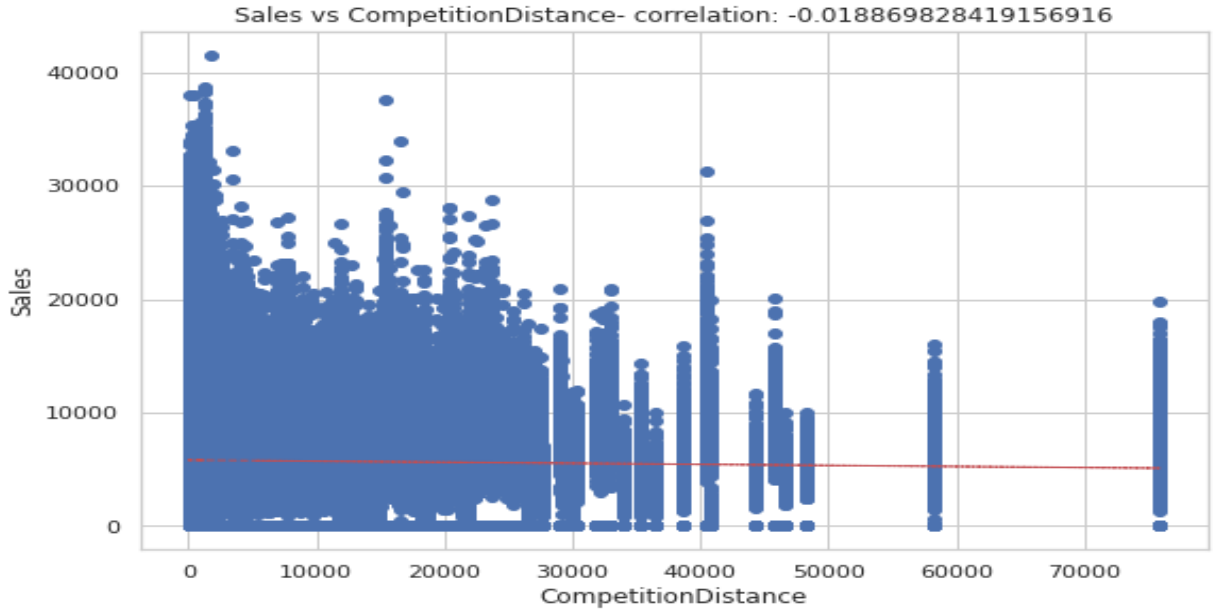
As we can see that the store who participating in promotion having more sales as compare to other. which means promotion is the key to attract the customer towards the store



EDA

Effect Of Competition Distance on Sales

Mostly stores were not that far from competitors and the stores were densely located near each other and surprisingly sales were higher when competition was nearer.

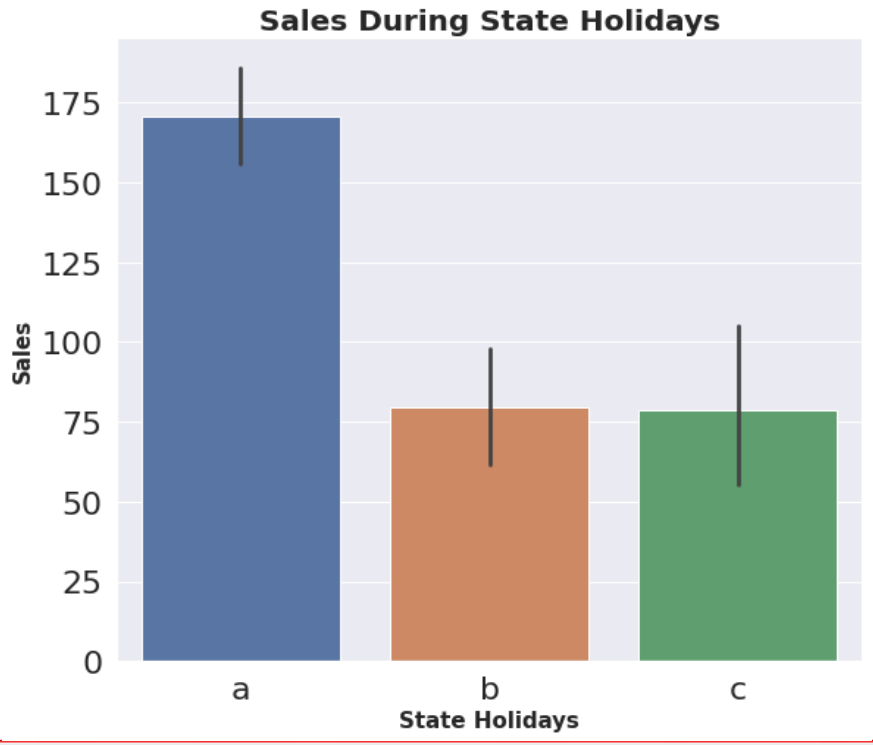


EDA

Sales During State Holidays

a = Public holiday, b = Easter holiday, c = Christmas

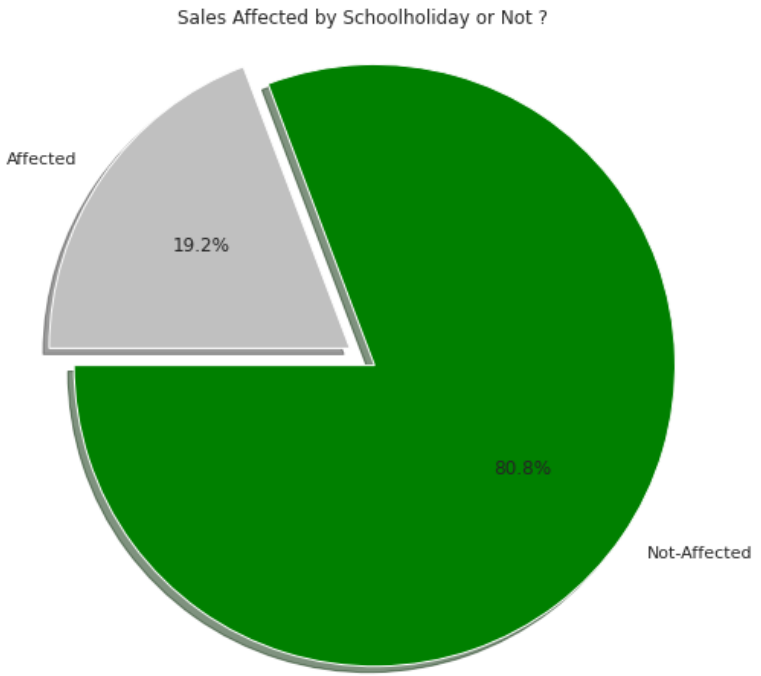
Stores has made more sales during Public holidays compared to Easter and Christmas holidays.



EDA

Impact Of School Holidays On Sales

19.20% of the total sales gets affected by the school holidays which also means that around 20% of the sales are oriented from the school students.

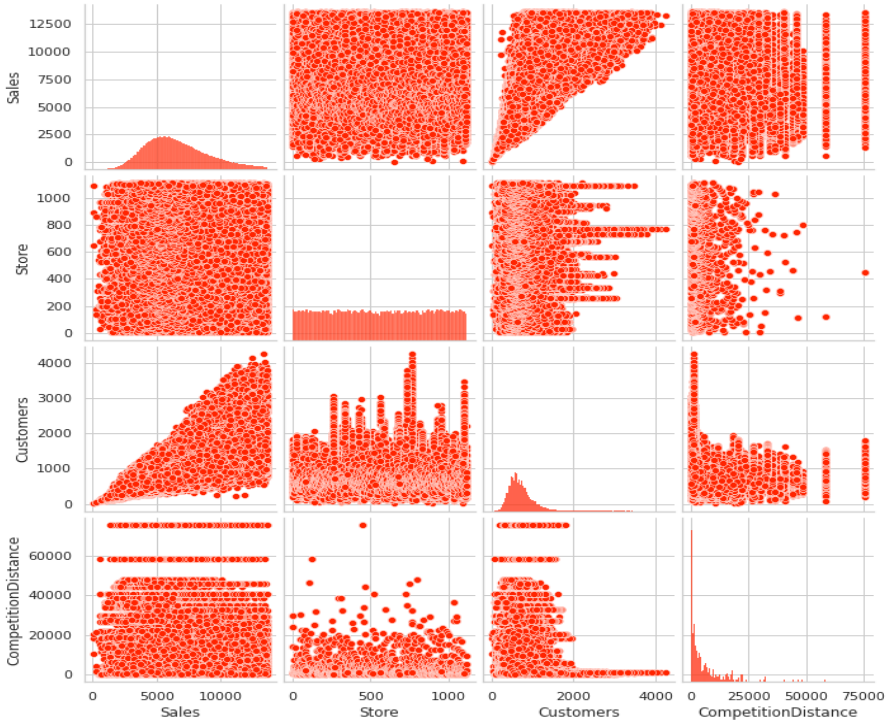


EDA

Linearity with Pair Plot

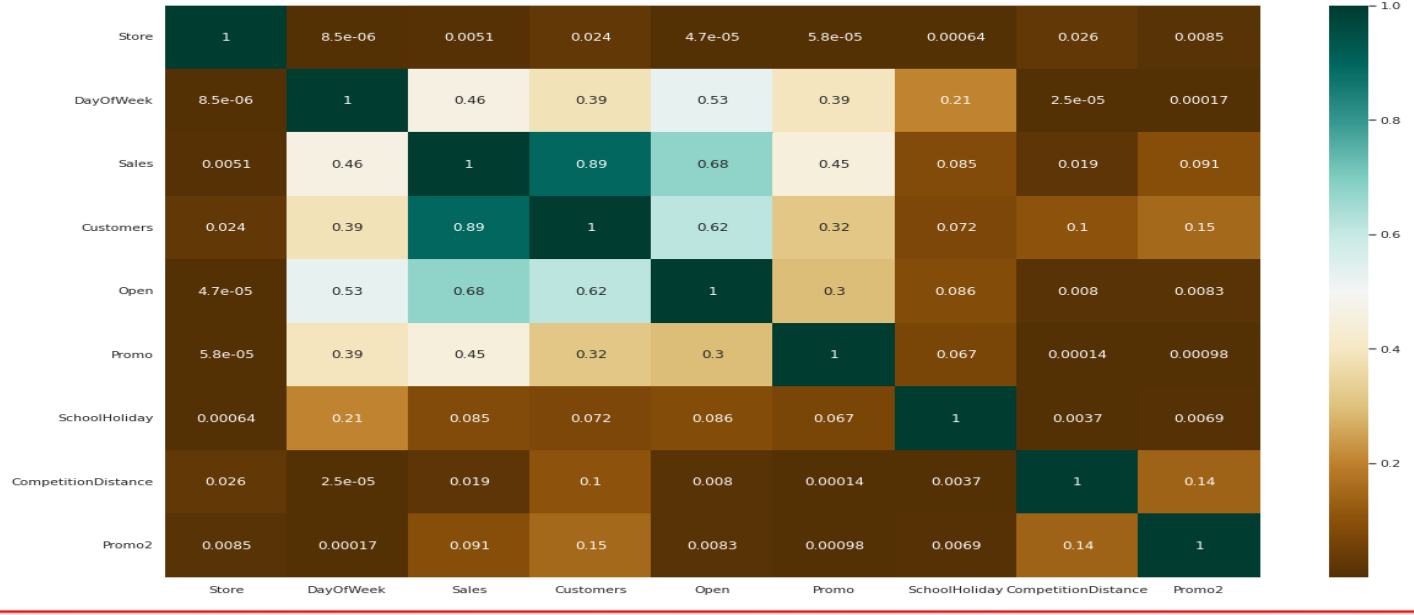
- Linear regression needs the relationship between the independent and dependent variables to be linear. So we used pair plot to check the relation of independent variables with the Sales variable.

- And we conclude that features like ‘Customers’, ‘CompetitionDistance’ and ‘Stores’ have a positive Relation.



Multicollinearity: We didn't find any correlation between independent variables but we found some correlation with our dependent feature which is a good sign for our model.

This plot shows that customer is highly co-related with sales.

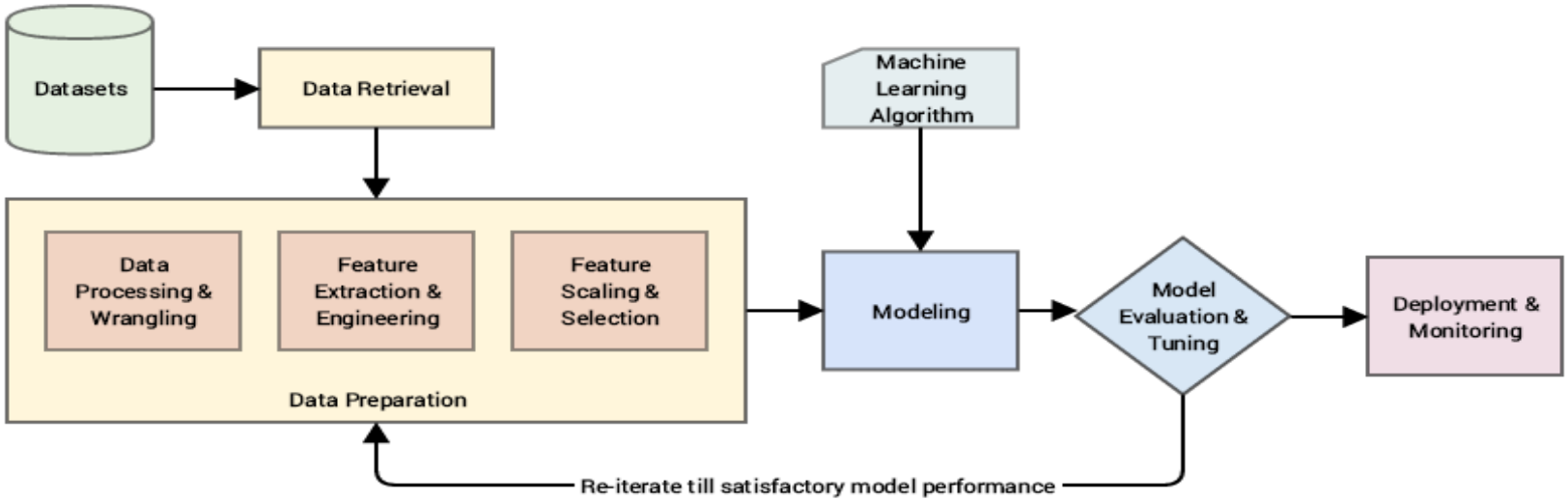


Features Transformation

- For column State Holiday we convert variable a, b, c into numerical variable 1 and existing variable 0 is kept as it is hence we got numerical feature having variables 0 and 1. It makes easy for computation in machine learning model.
- And we convert State Holiday column into integer Data type.
- Data Extraction: We have extracted Date, Year, Month from Date column for further analysis and then dropped the Date column.

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.



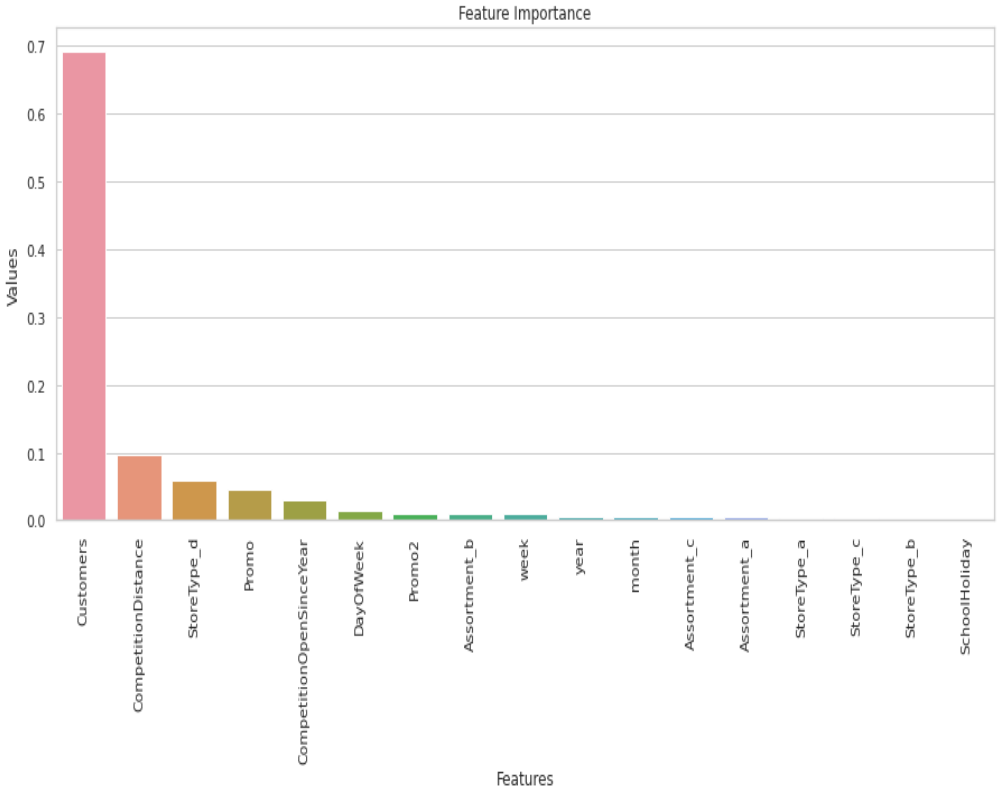
ML Model Performance

Looking at the various regression techniques we found out that 'Random Forest' have better model performance (Adjusted R2 : 0.9639) compared to other regression models.

Regression Techniques	MAE	MSE	RMSE	R2	Adjusted R2
Linear	866.335	1286832.8297	1134.3865	0.7797	0.7797
Lasso	866.2747	1286827.3908	1134.3841	0.7797	0.7797
Ridge	866.2762	1286826.1299	1134.3835	0.7797	0.7797
Elastic net Regress	874.1013	1331170.6124	1153.76	0.7721	0.7721
Decision Tree	445.8559	407405.8122	638.2834	0.9302	0.9302
Gradient Boosting	717.1779	871463.1338	933.5218	0.8508	0.8508
XGBoosting Regression	471.7540	393631.5969	627.4006	0.9326	0.9326
Random Forest	**326.8861**	**210858.0442**	**459.1928**	**0.9639**	**0.9639**

Feature Importance

After selecting our Random Forest Regression model we can see the importance of each features in our model prediction.



Challenges Faced

- Handling and understanding large amount of data.(1017209 number of records and 18 number of fields)
- Columns with improper data type and wrong values.
- Combining, creating and removing columns.
- Records containing more than 50% of nan values and replacing it with substitutes.
- Removing and replacing outliers from dependent and independent variables.
- Reducing skewness from the variables.
- Feature selections for ML Model.
- Converting columns with categorical variables to integer type and scaling numerical variables for regression models.
- Performing and choosing right kind of model.

Conclusions

- Store model 'b' have least number of stores in Rossmann yet it performed well and made more sales than other store models so it is advisable to increase the number of 'b' store model.
- Assortment level ‘Extra (1)' have the maximum number of stores in Rossmann yet it performed very badly but at the same time ‘Basic' and 'Extended' assortment level with less number of store had preformed extra ordinarily so it would be advisable to increase these assortment level.
- we can conclude that stores need more supply in between July to December stores should offer some discount in January to June to attract more customers.
- 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. Those products can be mainly be of basic necessities of a person's daily life.

Conclusions

- Average sales on weekdays was more as compared to weekends because promo's were provided to the customers during weekdays to increase the sales and not to weekends and reason might be that store use to remain close on Sundays.
- Sales during November and December month was higher compared to other months and that can be due to festive season in western European countries.
- School holidays also influenced the sales a lot as it can be observed that 17.8% of the sales gets affected by the school holidays which also means that around 17% of the sales are oriented from the school students.
- **Performing various regression techniques, we can observe that conclusion that 'Random Forest Regression' model have even higher performance (with R2 :0.9948) among the other models, as Random Forest Regression can handle large datasets efficiently and it's algorithm provides a higher level of accuracy in predicting outcomes over any other regression algorithm.**

AI

THANK YOU