Supervised ML – Regression

Capstone Project

On

Retail Sales Prediction

Team Members-

1. Raja Chowdhury
2. Saaquib Mustafa
3. Sandipan Das
4. Sahil Kolambkar

Rossmann

Dirk Rossmann GmbH, commonly referred to as **Rossmann**, is one of the largest drug store chains in Europe with around 56,200 employees. The company was founded in 1972 by Dirk Rossmann with its headquarters in **Burgwedel** near Hanover in Germany. The Rossmann family owns **60%** of the company. The Hong Kong based A.S. Watson Group owns 40% which was taken over from the Dutch Kruidvat in 2004.

The product range includes up to 21,700 items and can vary depending on the size of the shop and the location. In addition to drugstore goods with a focus on skin, hair, body, baby and health, pet food, a photo service and a wide range of natural foods and wines. Rossmann has 29 private brands with 4600 products (as of 2019).

In 2019 Rossmann had more than €10 billion turnover in **Germany, Poland, Hungary, the Czech Republic, Turkey, Albania, Kosovo and Spain**. In 2021, sales increased by 8.1 percent to **11.1 billion euros**. There are a total of 4,361 Rossmann branches, 2,231 of which are in Germany.

Points of Discussion

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

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.

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

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

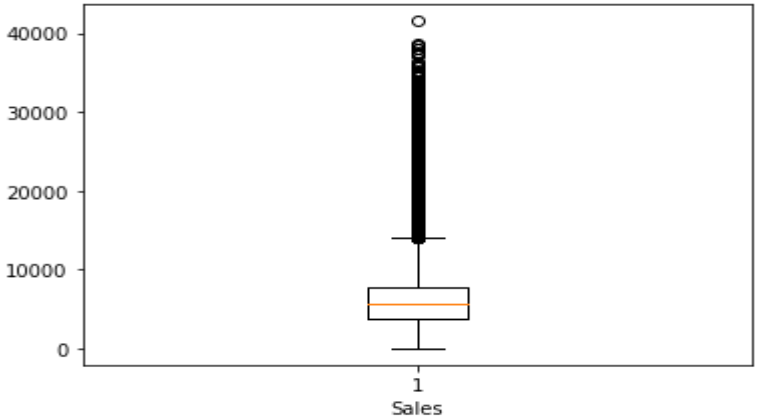
- **Merge Both Dataset:** We have merge both the available dataset
- **Data Extraction:** We have extracted Date, Year, Month, Day, Week, Week Of Year

  from Date column for further analysis and then dropped the Date column.
- **Combining and Creating Columns:**
  - We have created a new column called **'PromoOpen'** from existing columns to measure more accurate period in months from when the store is participating in Promo2.
  - We have created a new column called **'CompetitionOpen'** from existing columns to measure more accurate period in months from when the nearest competition has opened.
  - Then we replaced the negative values present in **'PromoOpen'** and **'CompetitionOpen'** with zero value.

- **Null Value Treatment:**
  - We have found out that whenever the store was not participating in Promo2, we had null values present in **'PromoOpen'** and **'PromoInterval'** columns. So we have imputed those null values with zero because logically when promo2 is zero then PromoOpen and PromoInterval should be zero as well.
  - We have replaced null values present in **'CompetitionDistance'** column with median as the CompetitionDistance's distribution was skewed towards right.
  - We have replaced null values present in **'CompetitionOpen'** column with mode as the CompetitionOpen column was made by combing the two categorical columns.
- **Changing Dtypes:**
  - '**CompetitionOpenSinceMonth'**, '**CompetitionOpenSinceYear**', '**Promo2SinceWeek**' and **'Promo2SinceYear**' columns are only using whole numbers and they have a discrete value, So we will change them from floats to integers.

➢ Handling Outliers:

- ❖ **'Sales'** and **'Customers'** columns were very important columns so outliers in these columns would have affected our prediction tremendously therefore we have removed them using z score method.
- ❖ **'CompetitiononDistance'**, **'CompetitionOpen'**, **'PromoOpen'** columns has huge number of outliers so we have replaced the outliers with different percentiles values using capping method.



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

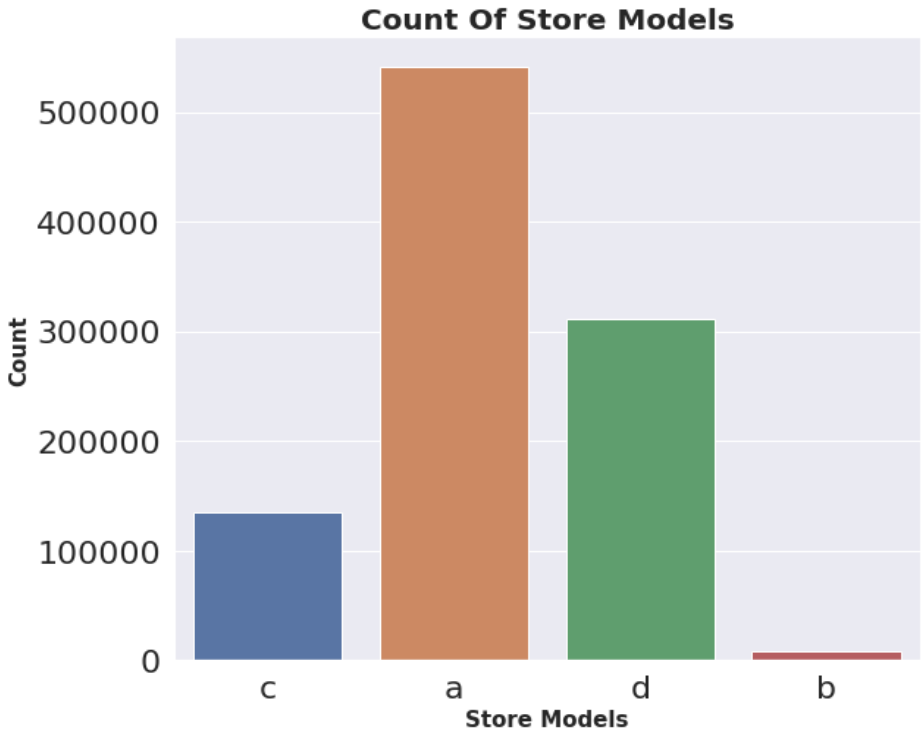
State Holidays

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

EDA

Store Models

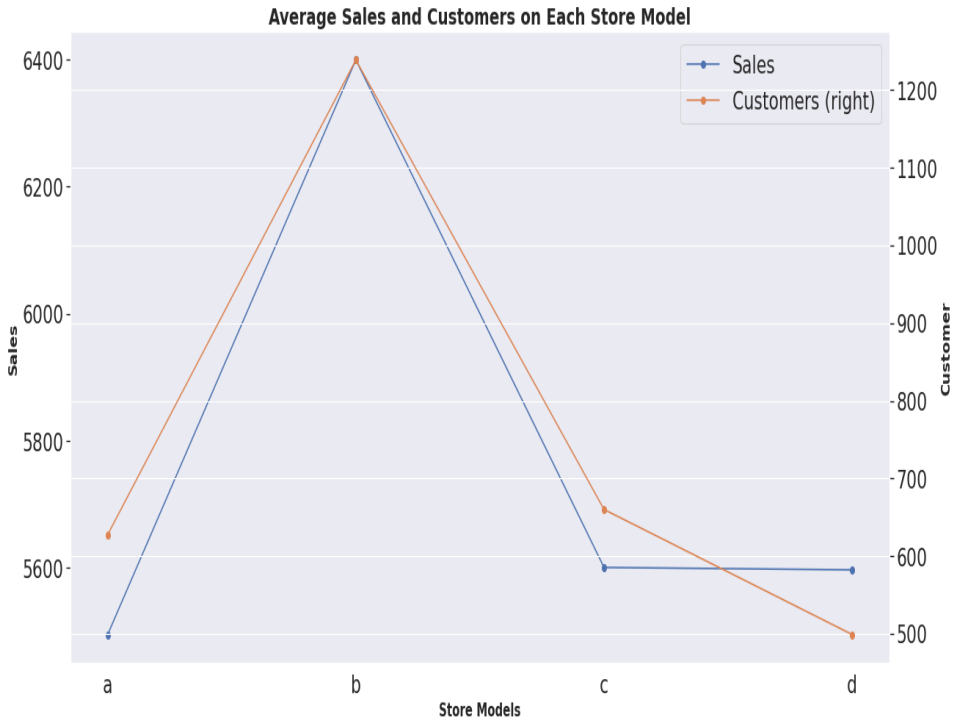
Store Model ’a’ have the maximum number of sales and store counts followed by ’d’ while Store Model 'b' have the least number of sales and store counts.



EDA

Average Sales and Customers of Store Models

Store model ‘b’ which have least number of store counts performed quite well on average sales and customers compared to other store models.

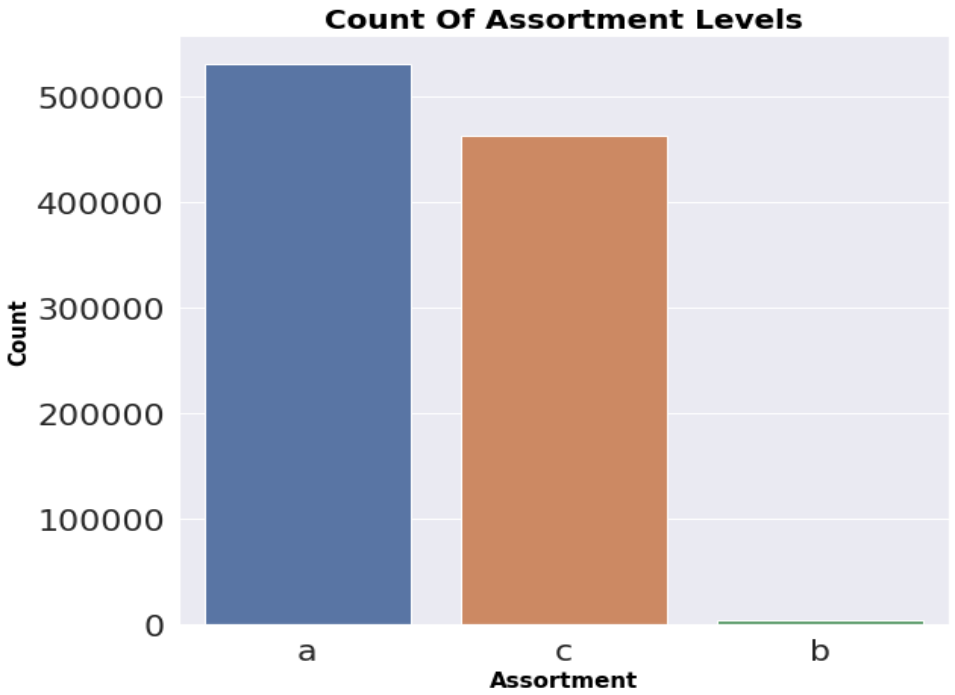


EDA

Assortment Levels

a = Basic, b = Extra, c = Extended

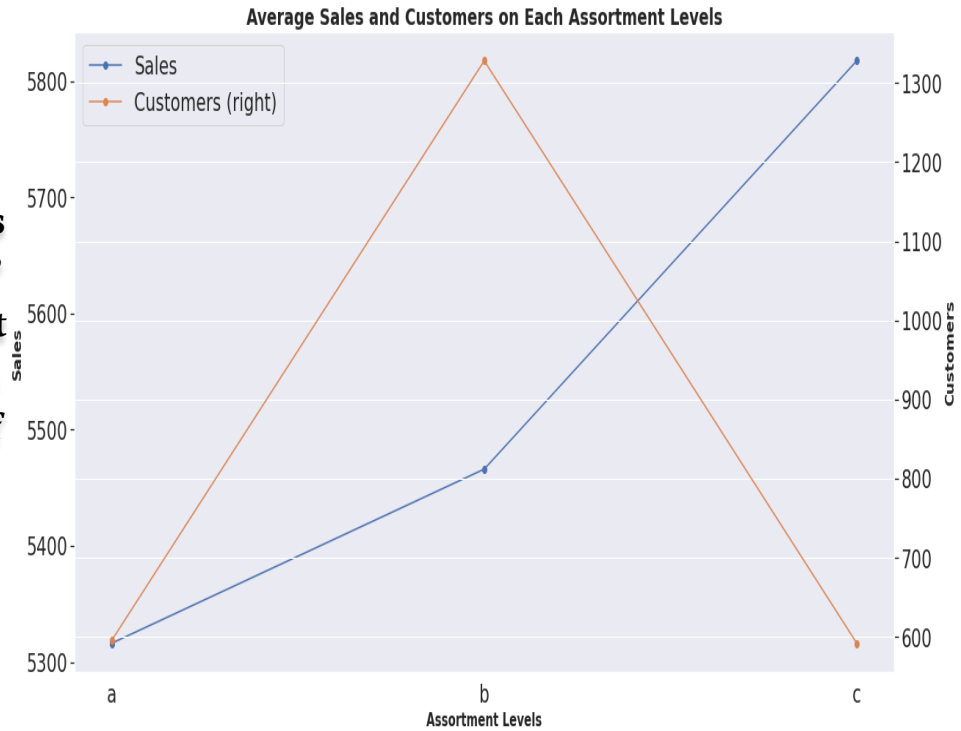
Basic Assortment level have maximum number of sales and store counts followed by Extended level while Extra Assortment have the least number of sales and store counts.



EDA

Average Sales and Customers of Assortment Levels

Assortment level 'b' with least store counts have perform quite well compared to 'a' while there is an another surprising fact that assortment level 'c' have maximum number of sales with the least number of customers.



EDA

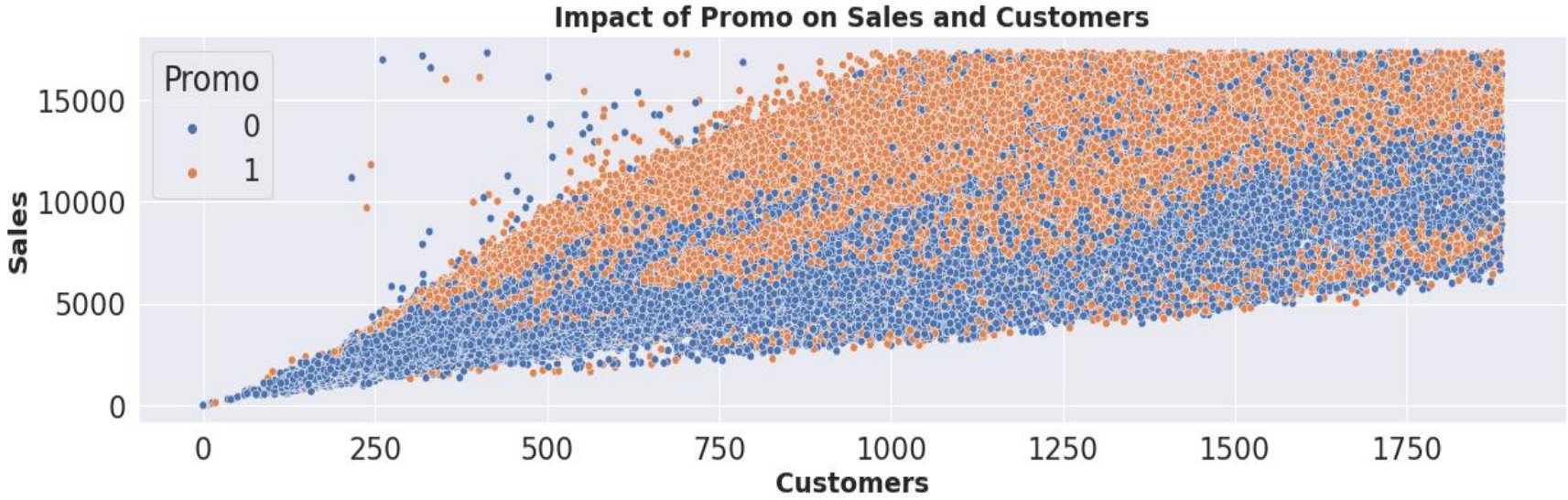
Sales In Different Stores And Assortment

Store Model ‘b’ has the maximum number of sales at all assortment level and assortment level ‘b’ is only available in store level ‘b’.

EDA

Impact Of Promo On Sales And Customers

There is a linear relationship between customers and sales and it is also noticeable that whenever the promo was open, stores has higher sales and customers compared to the similar period when promo was closed, which means promo had good impact on the business.

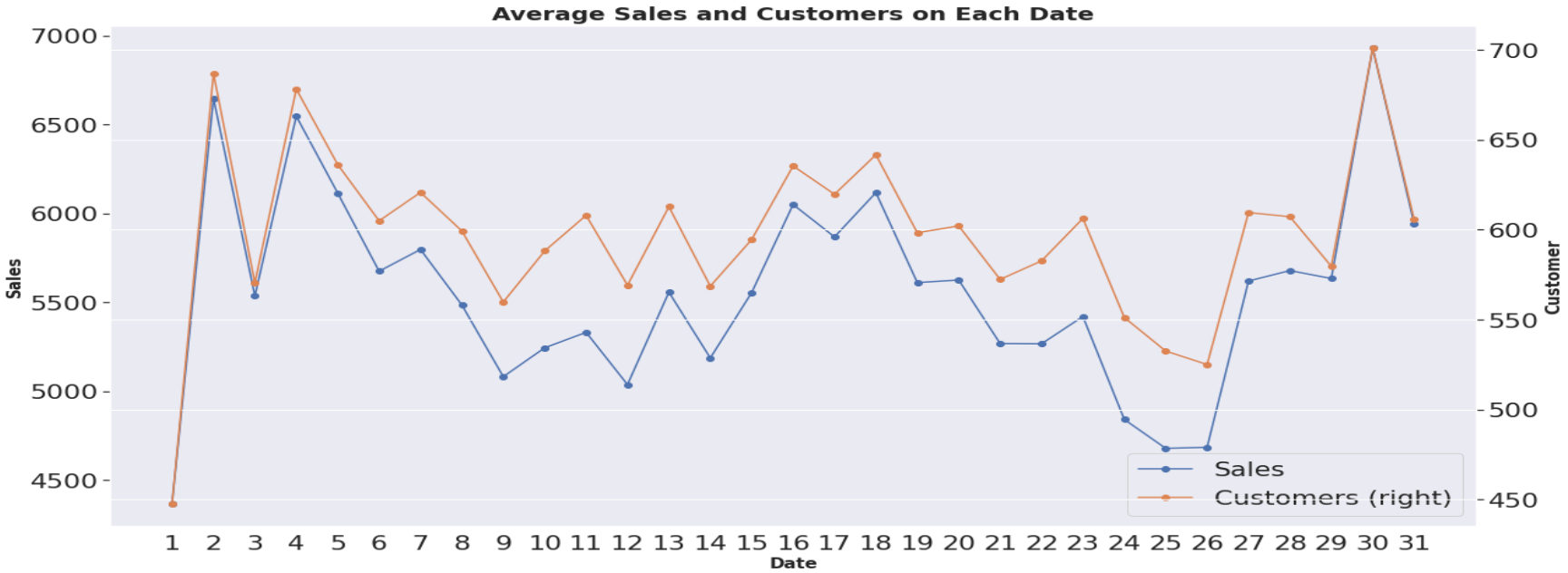


EDA

AI

Prices In Neighbourhood Groups

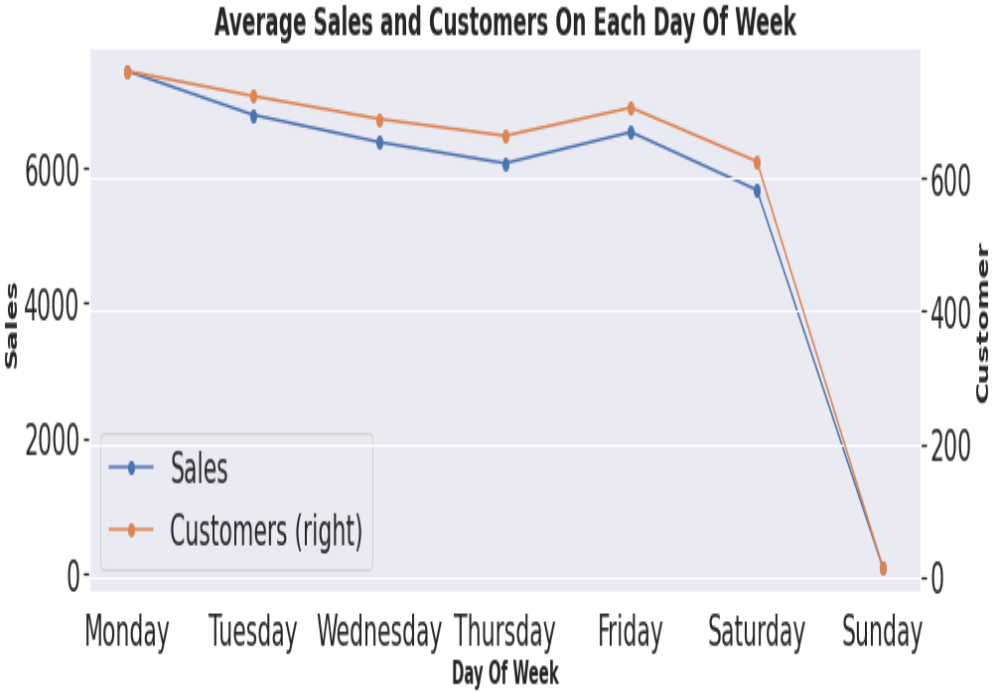
Sales and customers are highest on 30th followed by 2nd and 4th date of every month while sales and customers are lowest on the 1st date of every month followed by 25th and 26th date.



EDA

Average Sales And Customers On Each Day Of Week

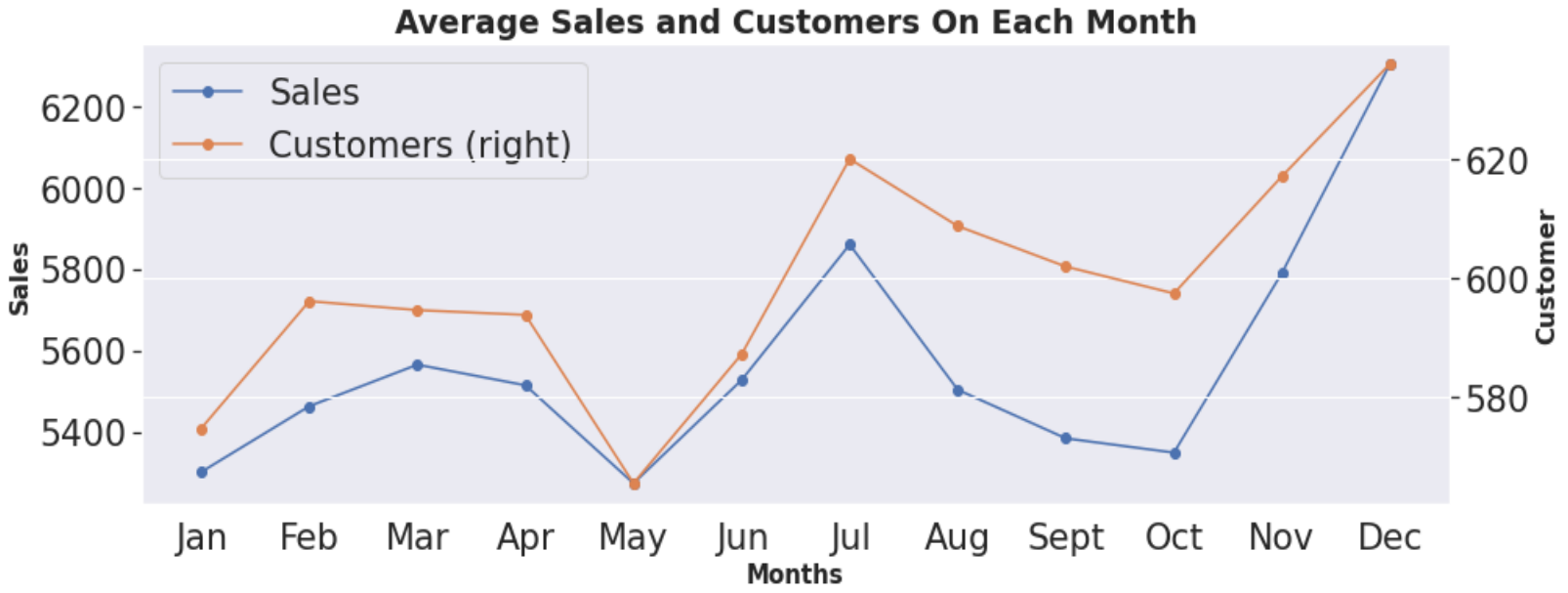
Sales and customers are at maximum on Mondays while sales and customers are nearly zero on Sundays because it seems like store use to remain closed on Sundays.



EDA

Average Sales And Customers On Each Month

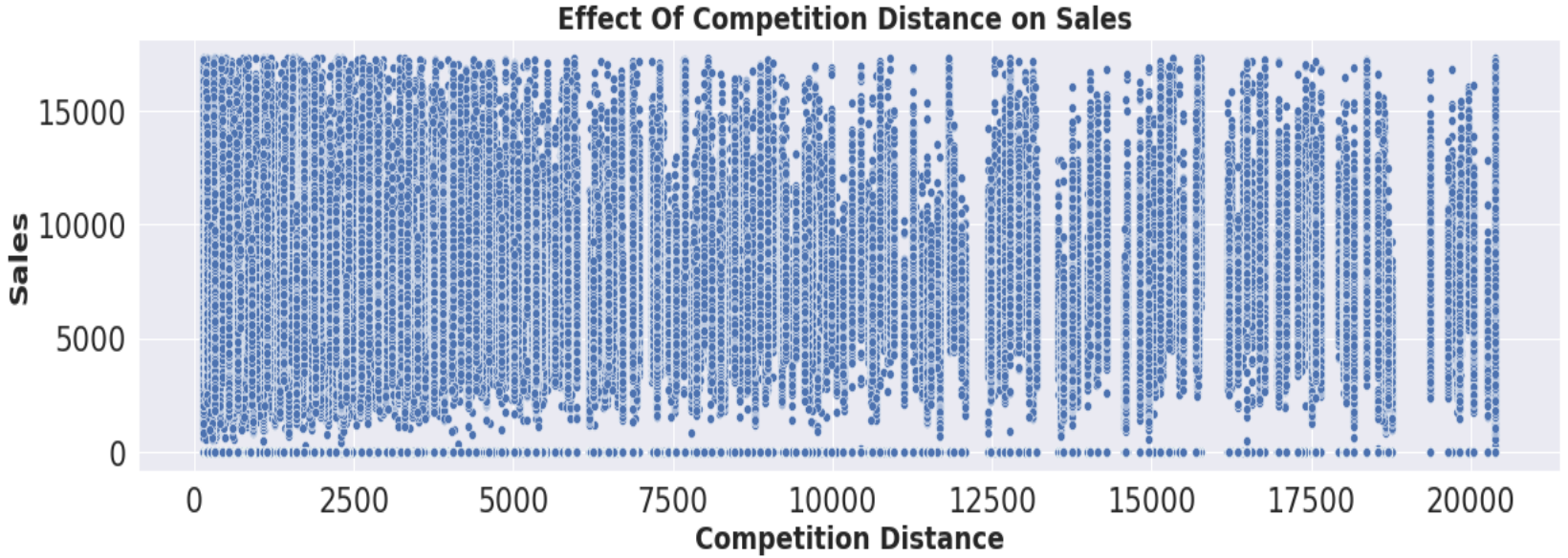
Sales and customers are at peak during November and December due to festive seasons like Christmas while sales and customers are at lowest during January and May or we may say these months to be the off-seasons.



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.

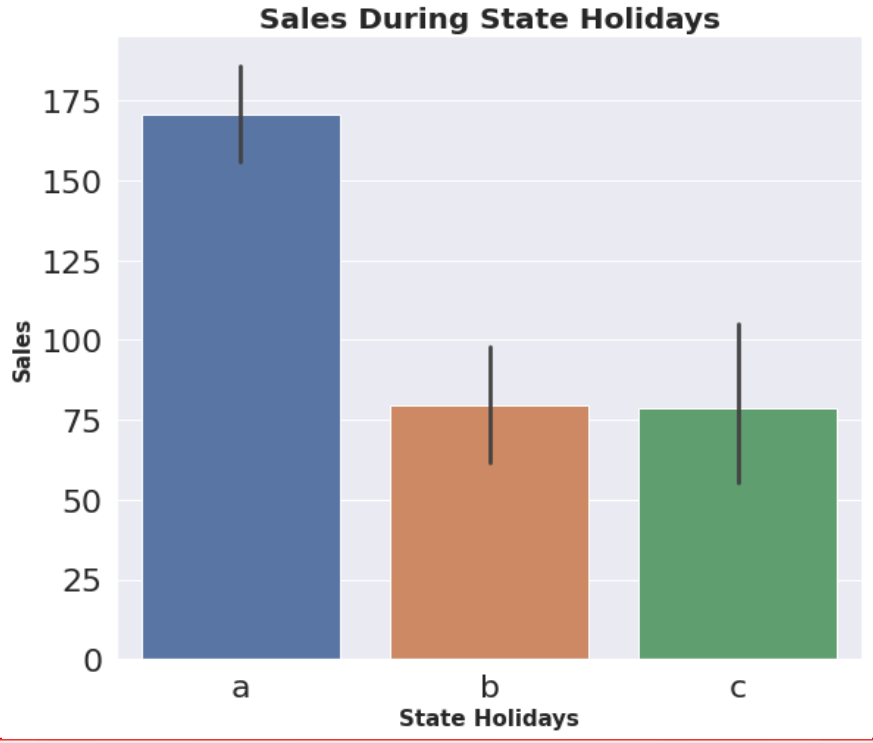


EDA

Sales During State Holidays

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

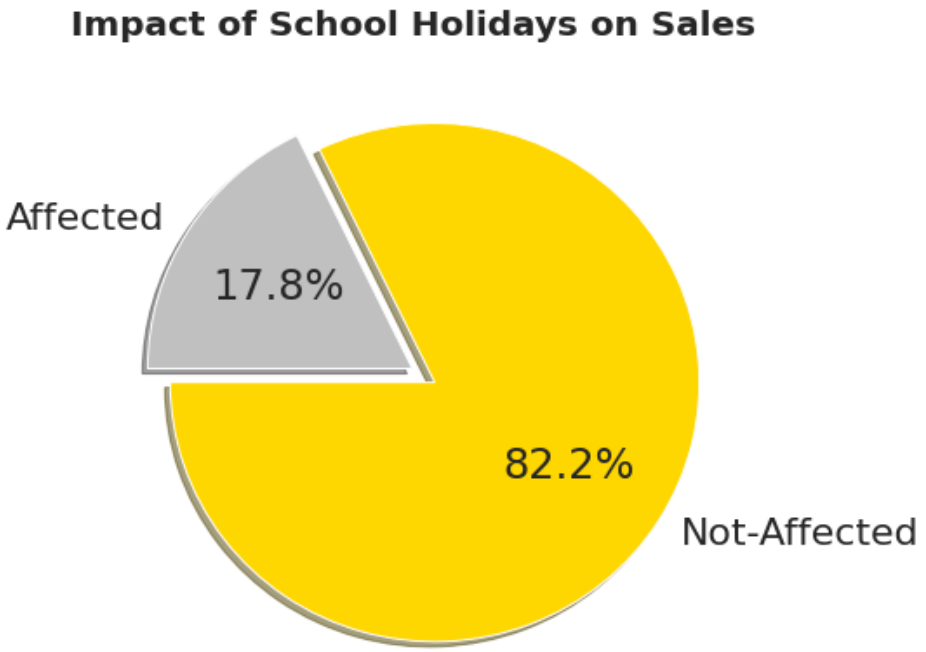
Stores has made more sales during Public holidays compared to Easter and Christmas holidays.



EDA

Impact Of School Holidays On Sales

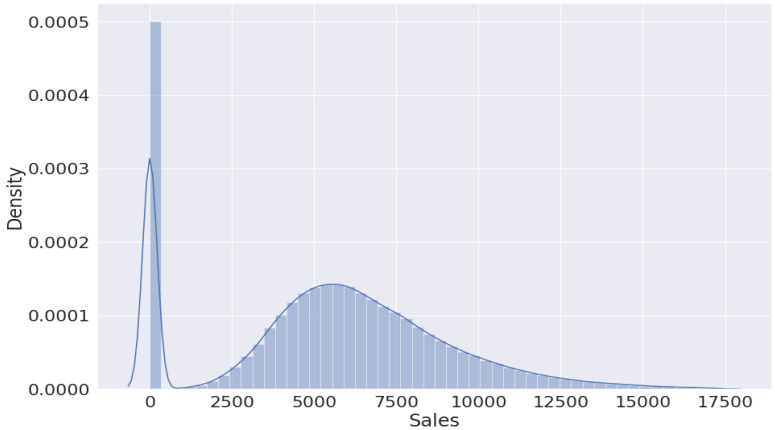
17.8% of the total sales gets affected by the school holidays which also means that around 17% of the sales are oriented from the school students



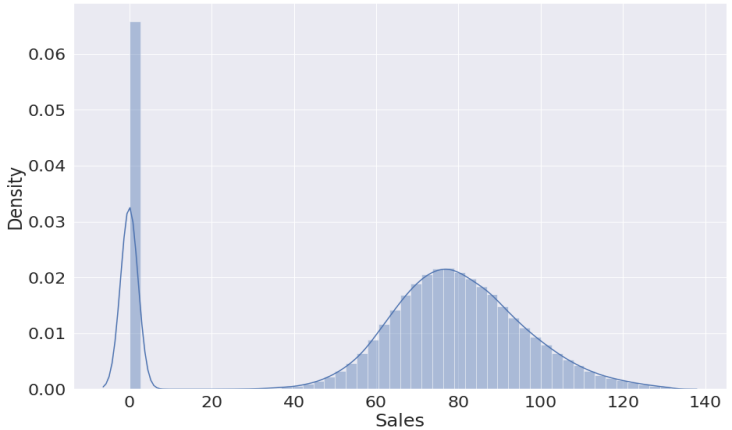
Feature Engineering

Before proceeding to ML Model we did some feature engineering to simplify and speed up data transformation while also enhancing model accuracy.

- **Dependent Variable Transformation:** We did sqrt transformation to remove the right skewness from the dependent variable.

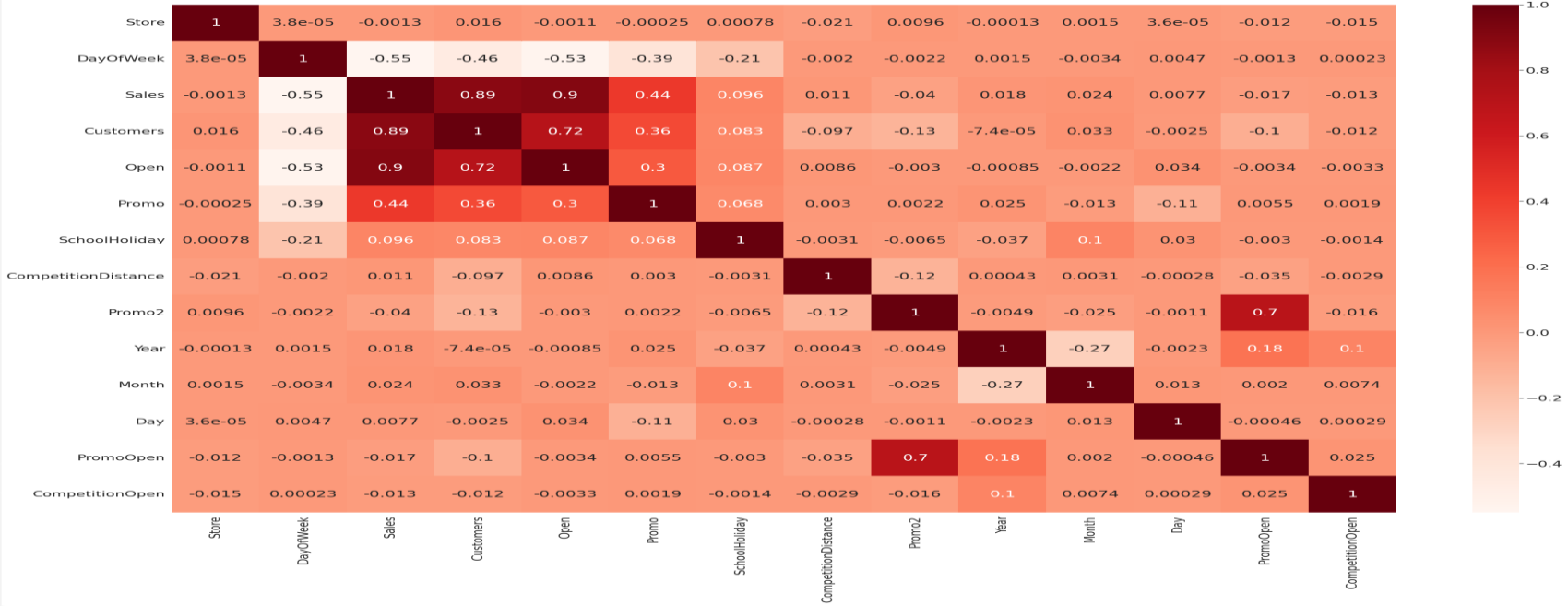


Before Sqrt Transformation



After Sqrt Transformation

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.



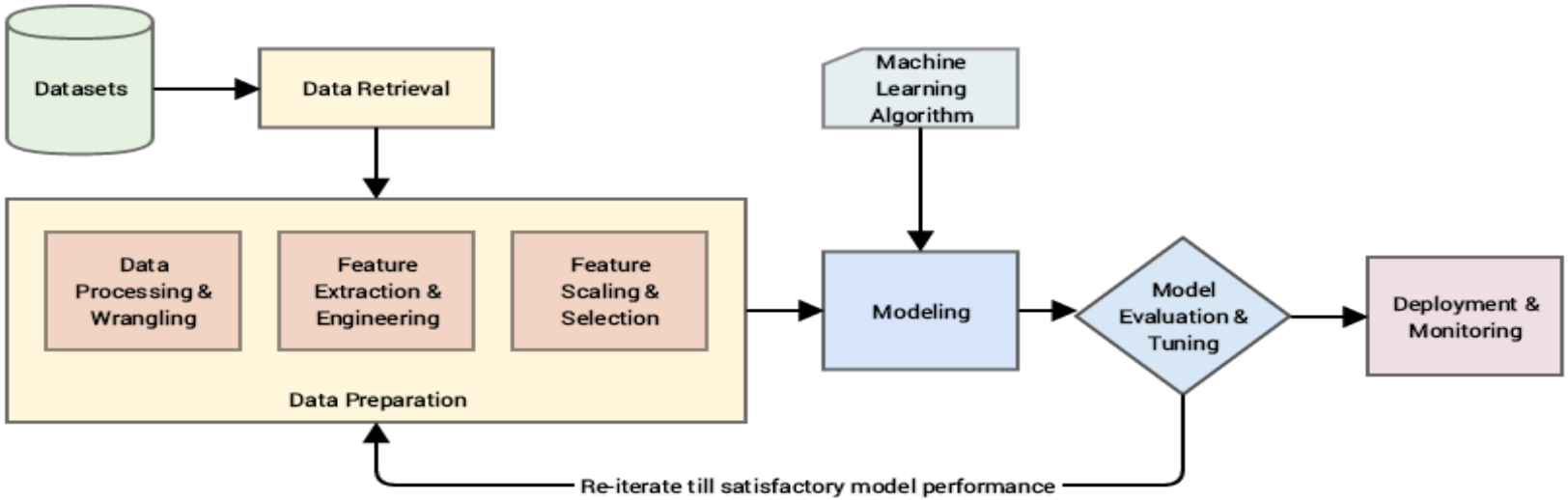
- **Feature Selection:** After analysis each columns we found out that ‘Store’ column have all unique values which won’t help me in any prediction so we dropped that column.

- **Scaling Numerical Feature:** We used MinMaxScaler to scale numerical variables **['Customers’, 'CompetitionDistance’, 'Year’, 'Month’, 'DayOfWeek’, 'Day’, 'CompetitionOpen’, 'PromoOpen’]** within a given range of 0 to 1 .

- **Dummification:** We used dummies to convert categorical variables **['StateHoliday’, 'StoreType’, 'Assortment’, ‘PromoInterval’]** into dummies variables of 1 and 0.

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 XGboost have better model performance (Adjusted R2 : 0.988407) compared to other regression models.

Regression Techniques	MAE	MSE	RMSE	RMPSE	R2	Adjusted R2
Linear	5.0754	42.0343	6.4833	0.0975	0.9628	0.9628
Lasso	7.4080	87.1560	9.3357	0.1404	0.9229	0.9229
Ridge	5.0754	42.0343	6.4833	0.0975	0.9628	0.9628
Decision Tree	2.3959	13.7608	3.7095	0.0558	0.9878	0.9878
Random Forest	4.1294	33.8483	5.8179	0.0875	0.9700	0.9700
Gradient Boosting	3.8355	27.1009	5.2058	0.0783	0.9760	0.9760
XGboost	2.6834	13.1049	3.6200	0.0544	0.9884	0.9884

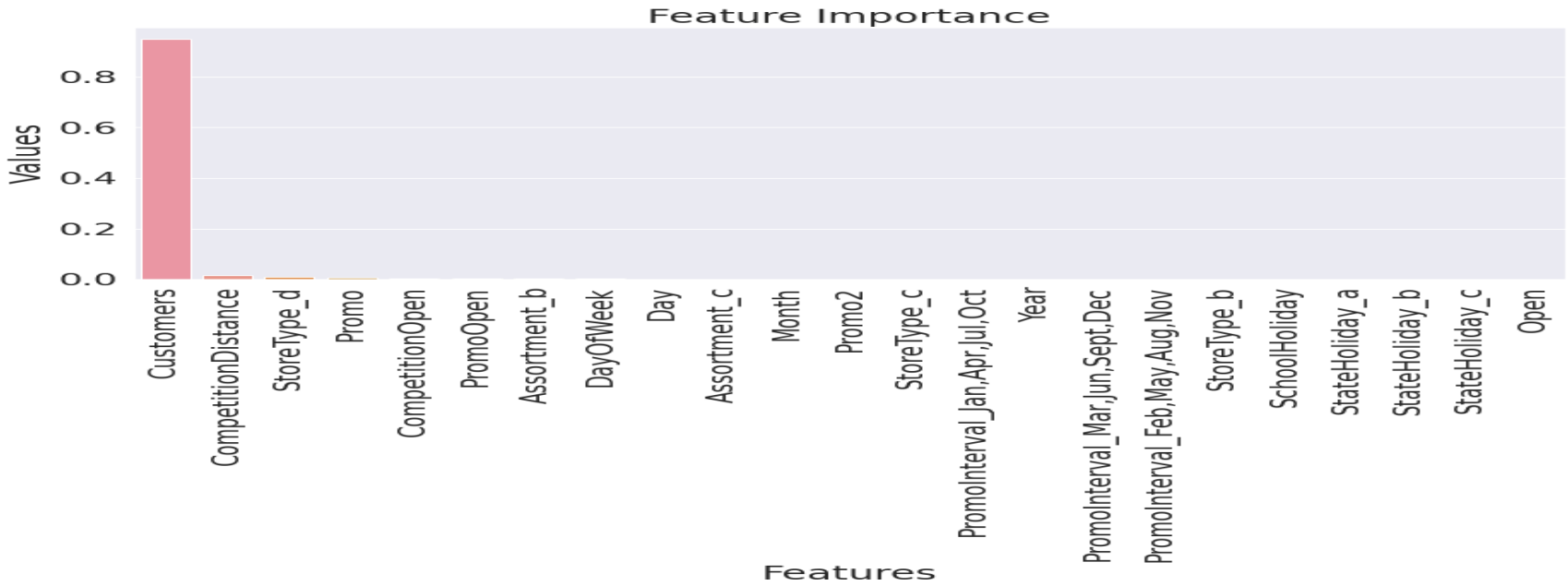
ML Model Performance

After using hyperparameter tuning on all regression models we came to our final conclusion that Random Forest Regression have performed much better with Adjusted R2 : 0.99409 as Random Forest Regression can handle large datasets efficiently and the random forest algorithm provides a higher level of accuracy in predicting outcomes over any other regression algorithm.

Hyper-Tuned Regression Techniques	MAE	MSE	RMSE	R2	Adjusted R2
Lasso	5.075486	42.034368	6.483392	0.96282	0.962816
Ridge	5.075483	42.034367	6.483392	0.96282	0.962816
Decision Tree	2.154891	11.09179	3.330434	0.990189	0.990188
Random Forest	**1.706587**	**6.680834**	**2.584731**	**0.994091**	**0.99409**
Gradient Boosting	3.308434	19.519283	4.418063	0.982735	0.982733
Xgboost	3.493811	22.214871	4.713265	0.980351	0.980349

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 'Basic' have the maximum number of stores in Rossmann yet it performed very badly but at the same time 'Extra' and 'Extended' assortment level with less number of store had preformed extra ordinarily so it would be advisable to increase these assortment level.
- Linear relationship have been found among customer, 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. Those products can be mainly be of basic necessities of a person's daily life.

Conclusions (cont.)

- 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 found that XGboost Regression model have the better performance (with R2 : 0.988409) but after applying hyperparameter tuning on all our models we finally came to the conclusion that Random Forest Regression model have even higher performance (with R2 :0.994091) 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.


END
CONCLUSION
OUTLINE

AI

THANK YOU