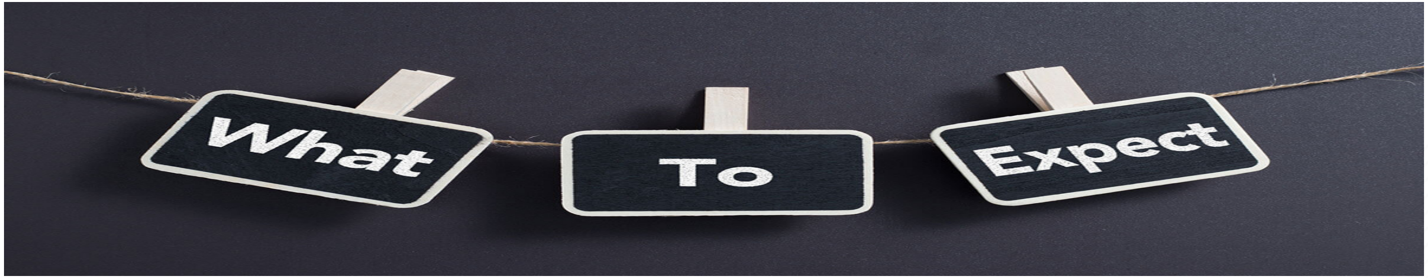
Capstone Project

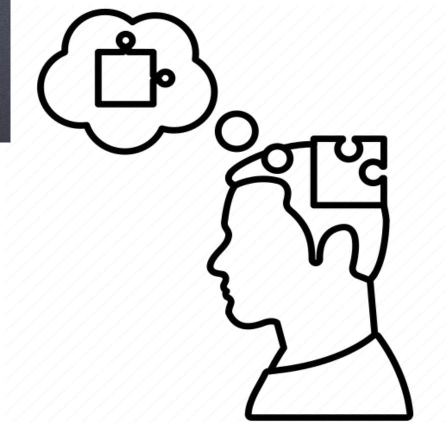
Retail Sales Prediction

Mind Benders Team Members

Abdul Aziz
Pooja Yadav
G M Sravya Sree
Abdullah Bin Mohammed



- Introduction to the Subject of Retail Sales Prediction.
- Problem Statement.
- Data Overview.
- Exploratory Data Analysis.
- Machine Learning Models
- Conclusions.
- Future Work.
- Reference



About Rossmann

- Rossmann is one of the largest drug store chains in Europe with annual revenues of almost $10 billion.
- Germany's first self-service drugstore.
- Rossmann operates over 3,000 drug stores in 7 European countries.
- The product range includes up to 21,700 items and can vary depending on the size of the shop and the location.

Problem Statement

- Analyzing the 2 years 7 months of historical sales data of 1,115 stores across Germany.
- Store Sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality
- Providing useful insights that help increase in productivity.
- To forecast the daily sale of individual 1115 Rossmann stores located across Germany, 6 weeks in advance.

Data Overview

Sr.No.	Data set	Variables	No. Of Variables	No. Of Observations
1	Rossmann Stores Data	Store, DayOfWeek, Date, Sales, Customers, Open, Promo, StateHoliday, SchoolHoliday	9	1017209
2	store	Store, StoreType, Assortment, Competition Distance, CompetitionOpenSinceMonth, CompetitionOpenSinceYear, Promo2, Promo2SinceWeek, Promo2SinceYear, PromoInterval	10	1115

Understanding Data

Sr. No.	Variables	Measurement Scale	Possible Values
1	Store	Nominal	1 to 1115
2	DayOfWeek	Nominal	1,2,3,4,5,6,7
3	Date	Interval	1/1/2013 to 7/31/2015
4	Sales	Ratio	0 to 41551
5	Customers	Ratio	0 - 7338
6	Open	Nominal	0(Closed) , 1(Open)
7	Promo	Nominal	0(No Promotion), 1(Offering Promotion)
8	State Holiday	Nominal	a: Public Holiday b: Easter Holiday c: Christmas Holiday d: None
9	School Holiday	Nominal	0(No), 1(Yes)

Continued...

10	Store Type	Nominal	a,b,c,d
11	Assortment	Nominal	a: basic b: Extra c: Extended
12	Competition Distance	Ratio	20 - 75860
13	Competition Open Since Month	Interval	1(Jan) to 12(Dec)
14	Competition Open Since Year	Interval	1990-2015
15	Promo 2	Nominal	0-1
16	Promo 2 Since Week	Nominal	1 - 50
17	Promo 2 Since Year	Nominal	2009 - 2015
18	Promo Interval	Ordinal	(Jan,Apr,Jul,Oct), (Feb,May,Aug,Nov), (Mar,Jun,Sept,Dec)

Data cleaning and feature engineering

```
Store                         0
StoreType                     0
Assortment                    0
CompetitionDistance           3
CompetitionOpenSinceMonth   354
CompetitionOpenSinceYear    354
Promo2                        0
Promo2SinceWeek             544
Promo2SinceYear             544
PromoInterval               544
dtype: int64
```

Checks the null values

Join the two datasets and imputed missing values

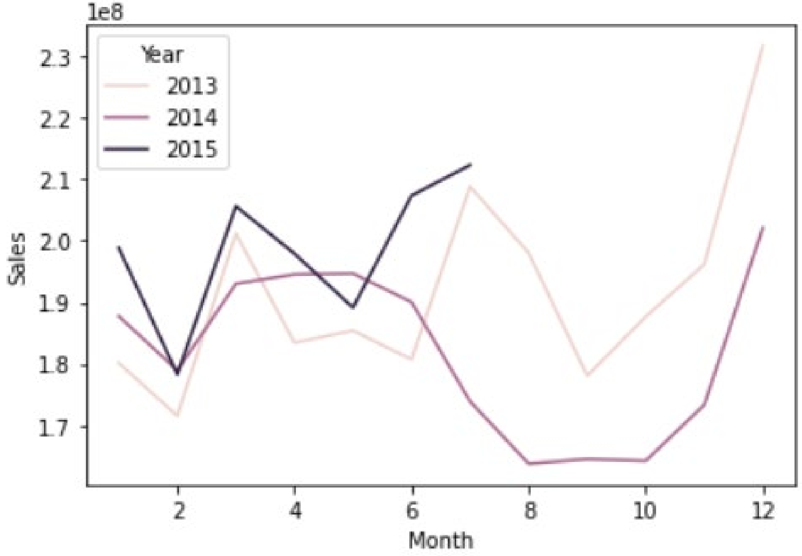
	Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	Year	Month	StoreType	Assortment	CompetitionDistance	Promo2	day_diff_comp	day_diff_promo
0	1115	2	0	0	0	0	a	1	2013	1	d	c	5350.0	1	0	212
1	504	2	0	0	0	0	a	1	2013	1	c	c	820.0	0	0	0
2	1016	2	0	0	0	0	a	1	2013	1	c	c	550.0	1	0	849
3	243	2	0	0	0	0	a	1	2013	1	a	a	310.0	1	0	-40
4	3	2	0	0	0	0	a	1	2013	1	a	a	14130.0	1	2223	632

```
Store                   0
Sales                   0
Customers               0
Open                    0
Promo                   0
SchoolHoliday           0
Year                    0
Month                   0
Assortment              0
CompetitionDistance     0
Promo2                  0
day_diff_comp           0
day_diff_promo          0
DayOfWeek_2             0
DayOfWeek_3             0
DayOfWeek_4             0
DayOfWeek_5             0
DayOfWeek_6             0
DayOfWeek_7             0
StateHoliday_0          0
StateHoliday_a          0
StateHoliday_b          0
StateHoliday_c          0
StoreType_b             0
StoreType_c             0
StoreType_d             0
dtype: int64
```

This is the final dataset and it can be used for regression model.

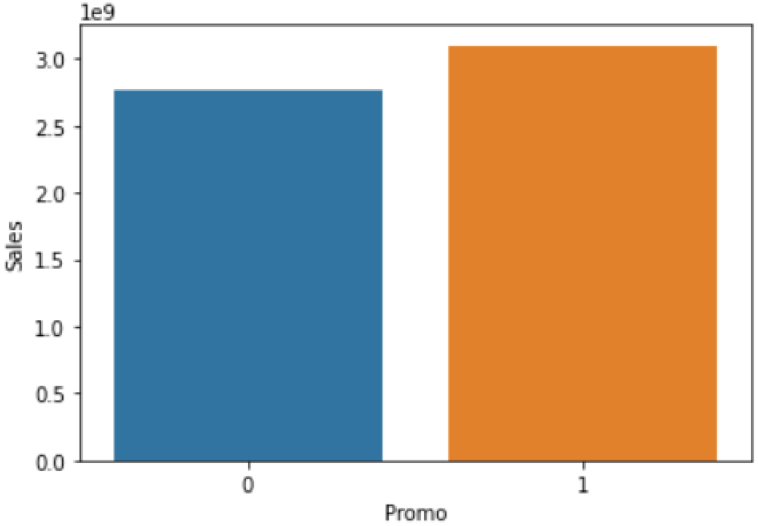
Exploratory Data Analysis

Sales across month



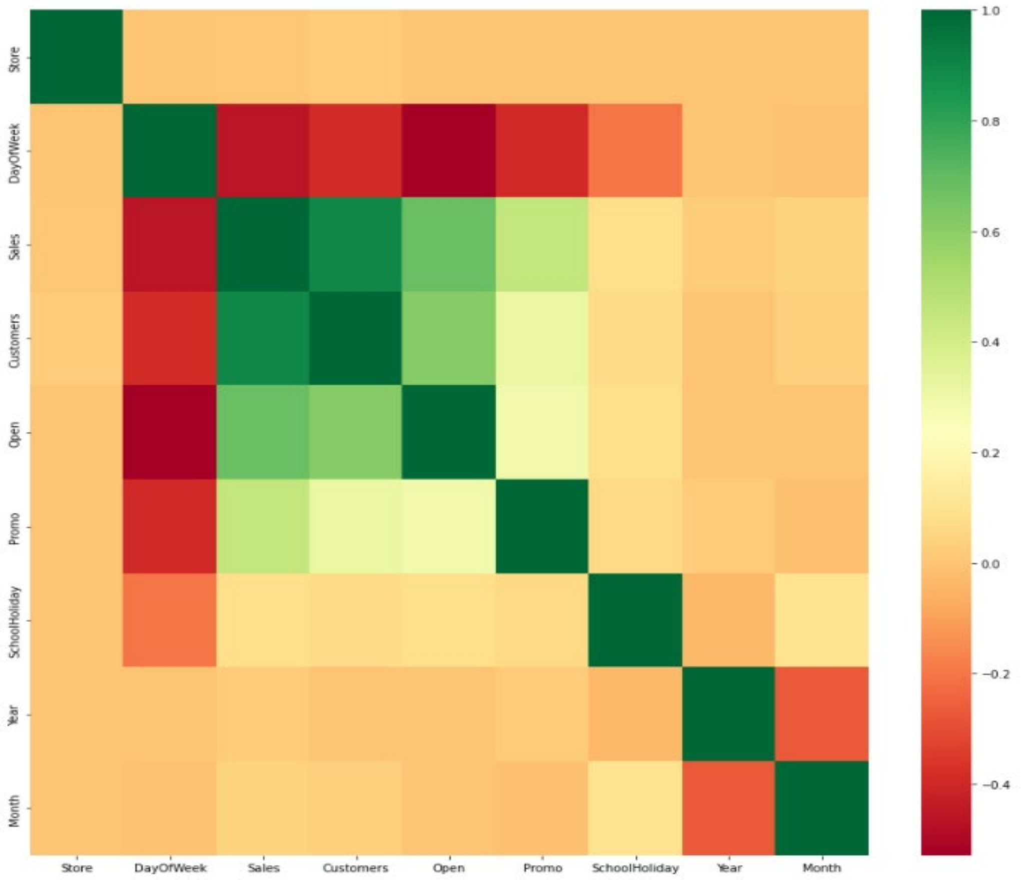
The sales is decreasing from January to February and then again increasingly sharply at the end of the year.

Effect of promo on sales



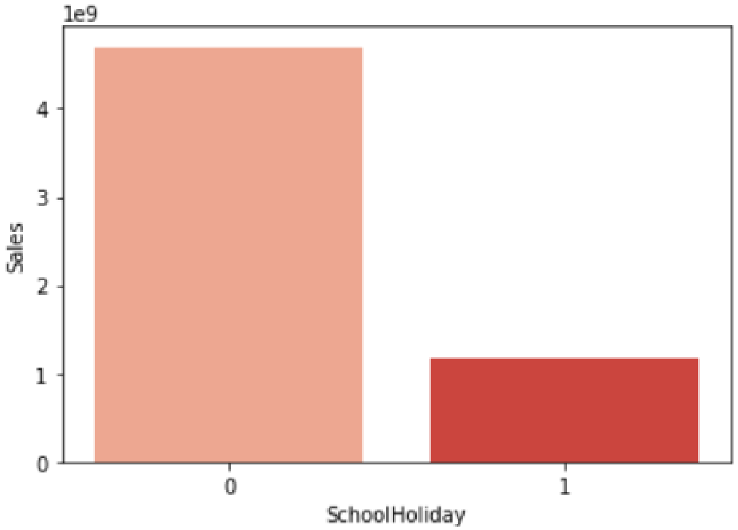
Promo had little effect on increasing the sales.

Correlation Matrix



Sales is highly correlated with number of customers.

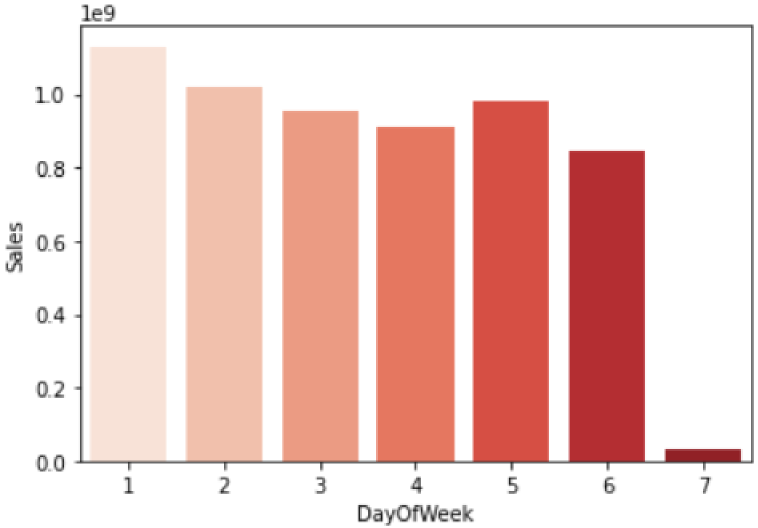
Sales vs School holiday



0 - No Holiday
1 - Holiday

Sales on days when Schools are open higher than School Holiday.

Sales across Day



Sales on seventh day of the week (Sunday) is extremely low compared to other dates.

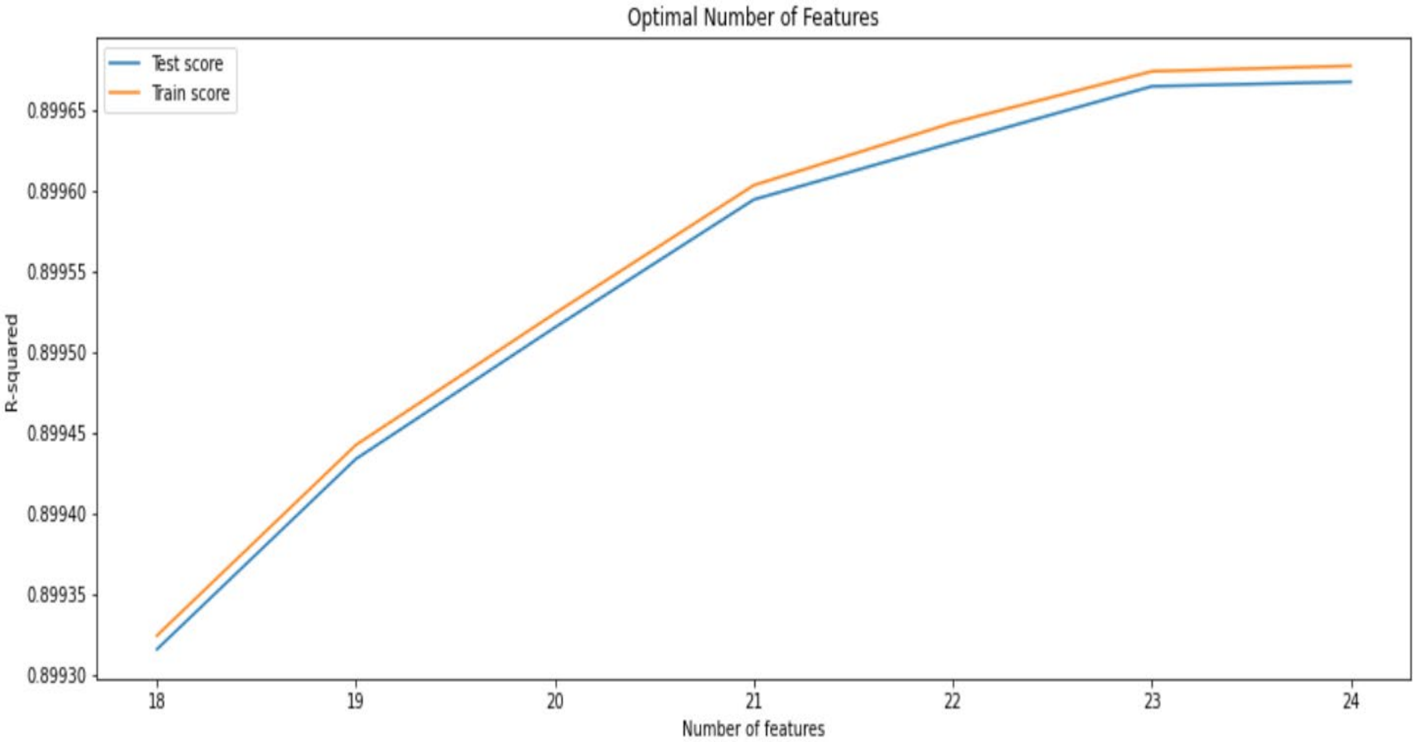
Machine Learning Models

Regression Models

- Linear Regression
- Random Forest
- XGBoost
- Decision Tree

Linear Regression

RFE with CV to find optimal number of features.



	mean_fit_time	params	mean_test_score	rank_test_score	mean_train_score
0	4.642537	{'n_features_to_select': 18}	0.899316	7	0.899324
1	4.094101	{'n_features_to_select': 19}	0.899434	6	0.899442
2	3.491326	{'n_features_to_select': 20}	0.899515	5	0.899524
3	2.900803	{'n_features_to_select': 21}	0.899594	4	0.899603
4	2.251261	{'n_features_to_select': 22}	0.899630	3	0.899642
5	1.575914	{'n_features_to_select': 23}	0.899664	2	0.899674
6	0.849403	{'n_features_to_select': 24}	0.899667	1	0.899677

Used cross-validation with RFE to obtain the number of features where we get best test score

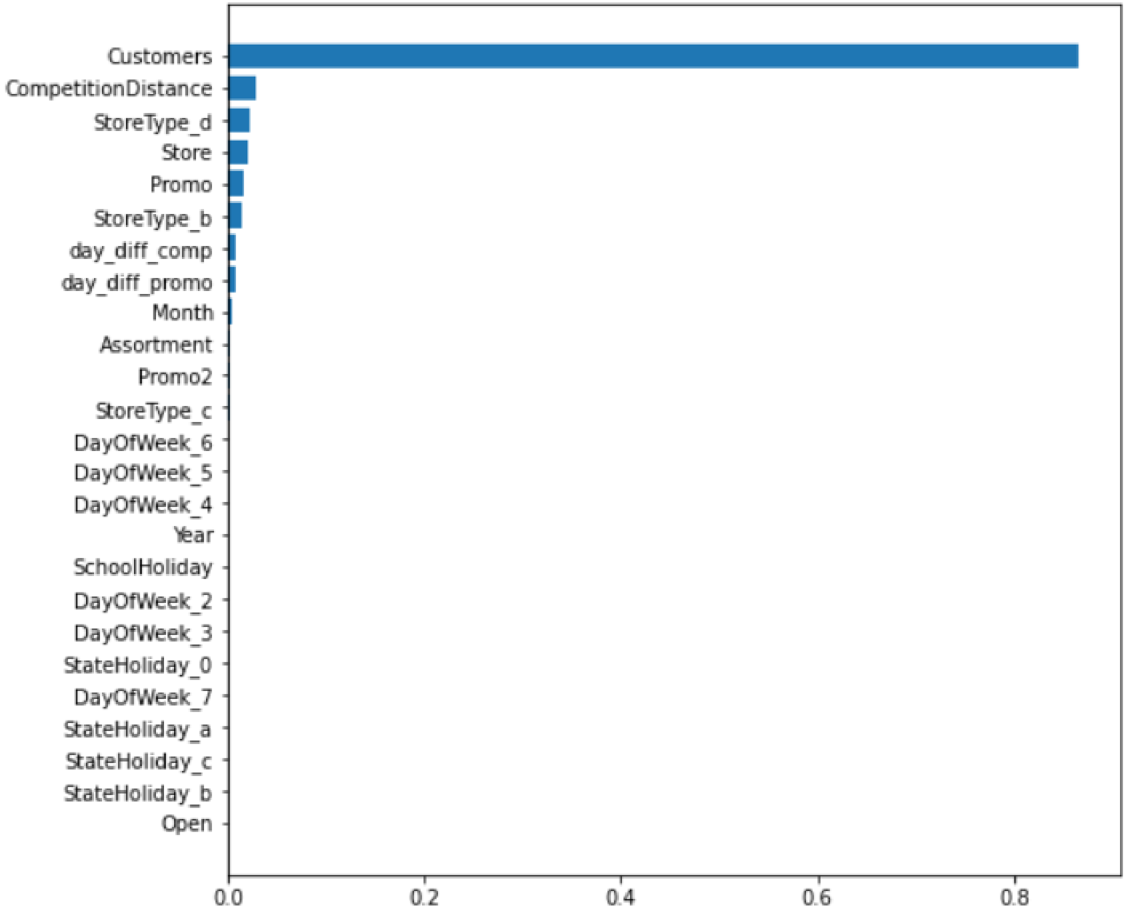
	Lin Reg Coeff	Lasso Reg Coeff	Ridge Reg Coeff
Customers	3343.297451	3341.968312	3342.421594
Open	405.206604	415.435482	404.851784
Promo	595.450396	595.396951	595.492138
SchoolHoliday	37.957593	37.179206	37.954846
Year	56.482944	54.952577	56.479233
Month	87.867831	86.950113	87.893193
Assortment	115.880008	115.094676	115.937216
CompetitionDistance	154.464806	153.410270	154.384150
Promo2	47.450532	46.284727	47.342093
day_diff_comp	24.445705	23.426389	24.436373
day_diff_promo	88.803968	88.680075	88.802423
DayOfWeek_2	-169.999621	-163.698400	-169.932955
DayOfWeek_3	-219.127670	-212.838688	-219.075637
DayOfWeek_4	-241.685269	-235.469605	-241.621360
DayOfWeek_5	-191.446003	-185.234256	-191.367931
DayOfWeek_6	-74.284518	-68.237714	-74.282836
DayOfWeek_7	-103.911605	-89.282687	-104.637246
StateHoliday_0	54.093962	52.876801	54.096654
StateHoliday_a	-13.719321	-9.755378	-14.018442
StateHoliday_b	-38.932217	-35.826989	-39.089559
StateHoliday_c	18.153901	18.716387	17.988143
StoreType_b	-646.121597	-645.318107	-645.737283
StoreType_c	-44.276995	-43.329812	-44.279582
StoreType_d	418.747280	418.210917	418.568329

- Comparison of coefficient for Linear and regularized models.
- We observe that the number of customers, Promo and Store type d and Store type c significantly affect the sales.

	Linear Regr	Lasso Regr	Ridge Regr
R2 Score	0.894335	0.894230	0.89433
Adj R2 Score	0.894320	0.894215	0.89432

Score comparison for linear and regularized models

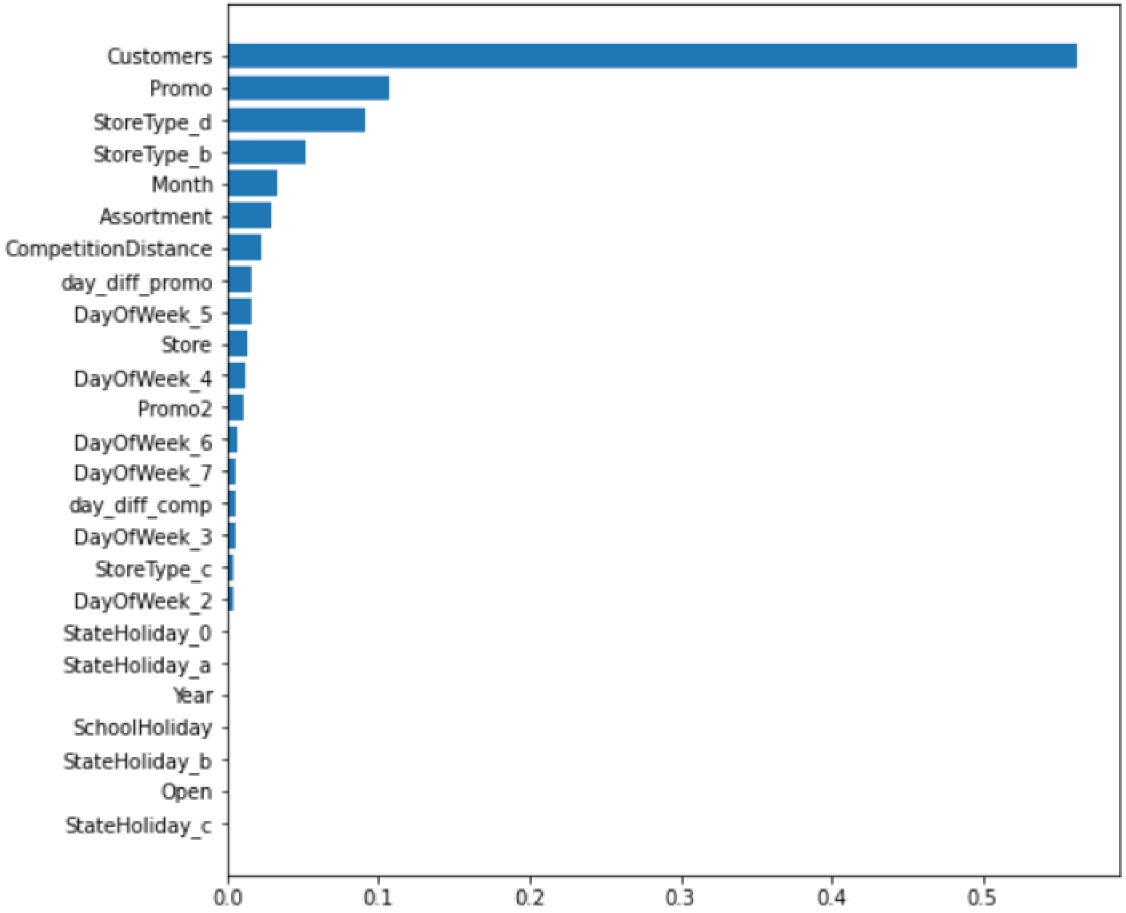
Random Forest



Feature importance for customers is high compared to others followed by competition Distance and store type d.

R-square value using Random Forest: 0.978

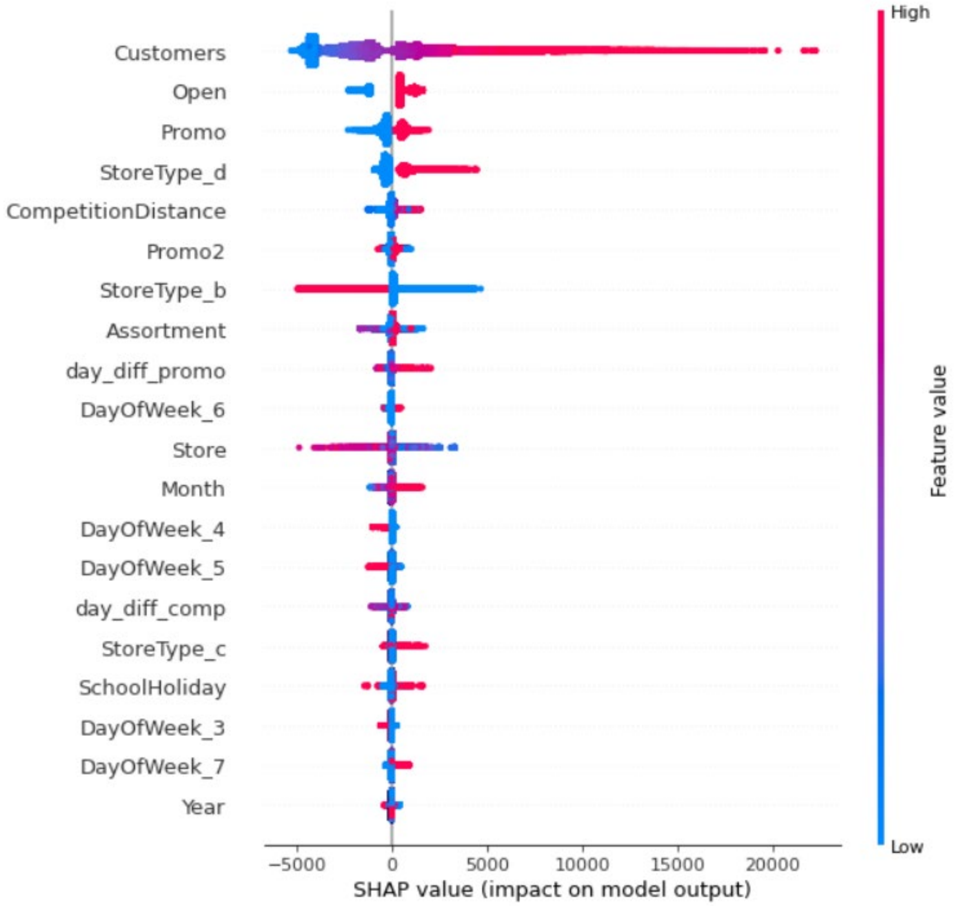
XGBOOST



Feature importance for customers is high compared to others followed by promo and store type d.

R-square value using XGB: 0.926

Decision Tree



- Feature importance for customers is high compared to others followed by Open and Promo.
- when customers are high it is evident that sales are high

R-square value using Decision Tree: 0.958672

Conclusion:

	Linear Regr	Decision Tree Regr	XGB Regr	Random Forest
R2 Score	0.894335	0.958672	0.925541	0.978277
Adj R2 Score	0.894320	0.958667	0.925532	0.978274

- The table above shows the R-square value obtained for the four models i.e Linear Regression, Decision Tree, Extra Gradient Boost (XGB) and Random Forest. Based on the scores, we can select a model depending on the business requirement. If we want a simple model that explains the variation of target with the features we can go for Linear regression or Decision tree. If we want a better accuracy we can go for complex models like XGB and Random Forest.

Challenges:

- Handling large amount of sales data(10,17,210 observation on 13 variables.)
- Some 180 stores were closed for 6 months. Unable to fill the gap of sales for those stores.
- Prediction of sales for individual stores (out of 1115) and most of stores have different pattern of sales.

  A single model cannot fit to all stores.

Learnings:

1) Exploring large datasets using visualisation tools.

2) Learn the application of Linear Regression, Random Forest, KNN Regression, Decision Tree.

Scope of Improvement :

- Applied only Four algorithms i.e, Random forest, XGB Regression, Decision Trees, and Linear Regression. So there are scope for applying more algorithms like SVM, Time Series analysis, KNN.
- For future work we can extract more features from the dataset or we can test with additional hyper parameter values to find the most optimal ones.

References

- Kaggle competition
- Analytics vidhya

Thank You!!