Capstone Project

Retail Sales Analysis

by

Mayank Sawant

Data Science Practitioner

AlmaBetter, Bengaluru

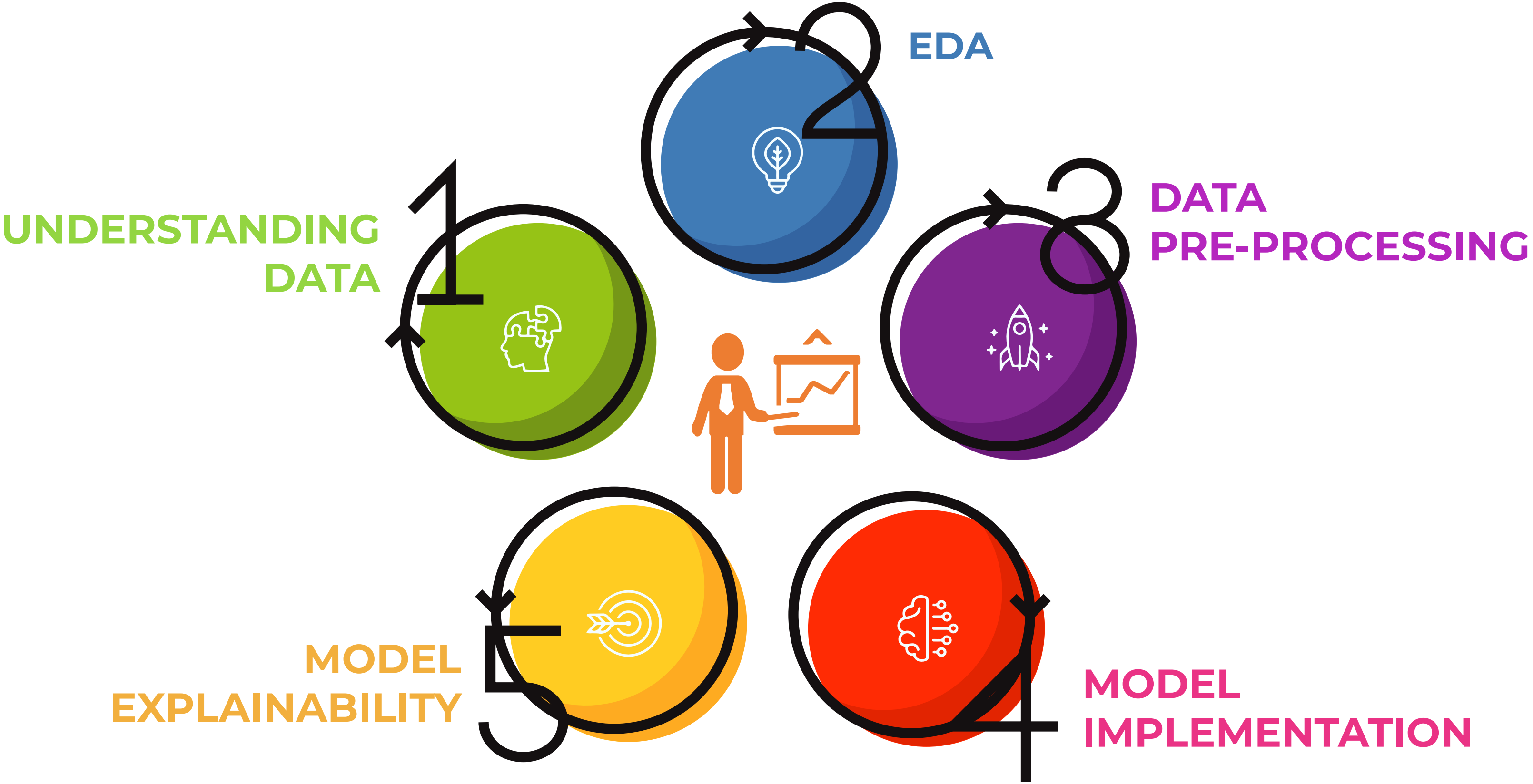
POINTS FOR DISCUSSION

1. Problem Statement
2. Model Architecture
3. Data Summary
4. EDA
5. Model Implementation

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.

Predicting Sales of a Major Store



1 UNDERSTANDING DATA

2 EDA

3 DATA PRE-PROCESSING

4 MODEL IMPLEMENTATION

5 MODEL EXPLAINABILITY

MODEL ARCHITECTURE

DATASET

Data Shape:
Sales Data - (1017209, 9)
Stores Data - (1115,10)

Store
Store id

Id
unique id

Sales
Sales made for the day

Customers
Footfall for the day

Open
Opened or closed

StateHoliday
Stateholiday or not

Assortment
Types of Assortment

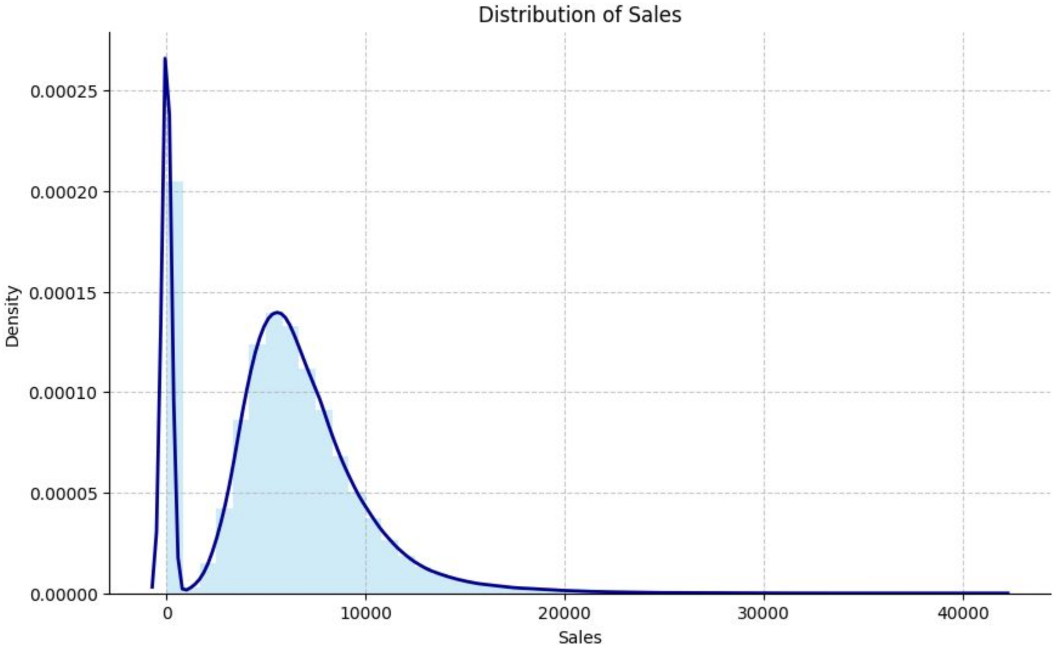
SchoolHoliday
Schoolholiday or not

Promo
Store running promotion or not

StoreType
Types of stores

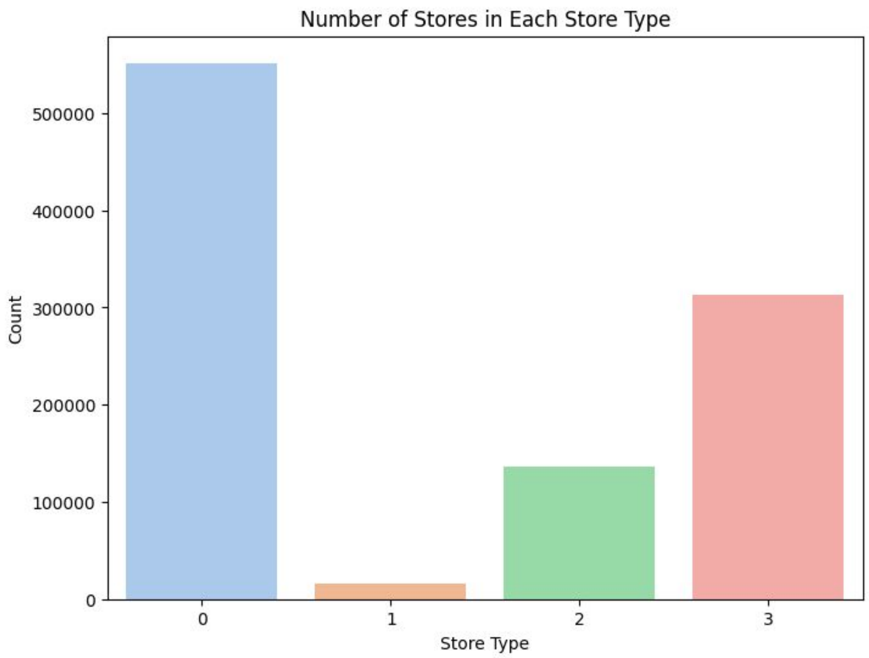
Promo2
Store running consecutive promotion or not

DISTRIBUTION OF SALES



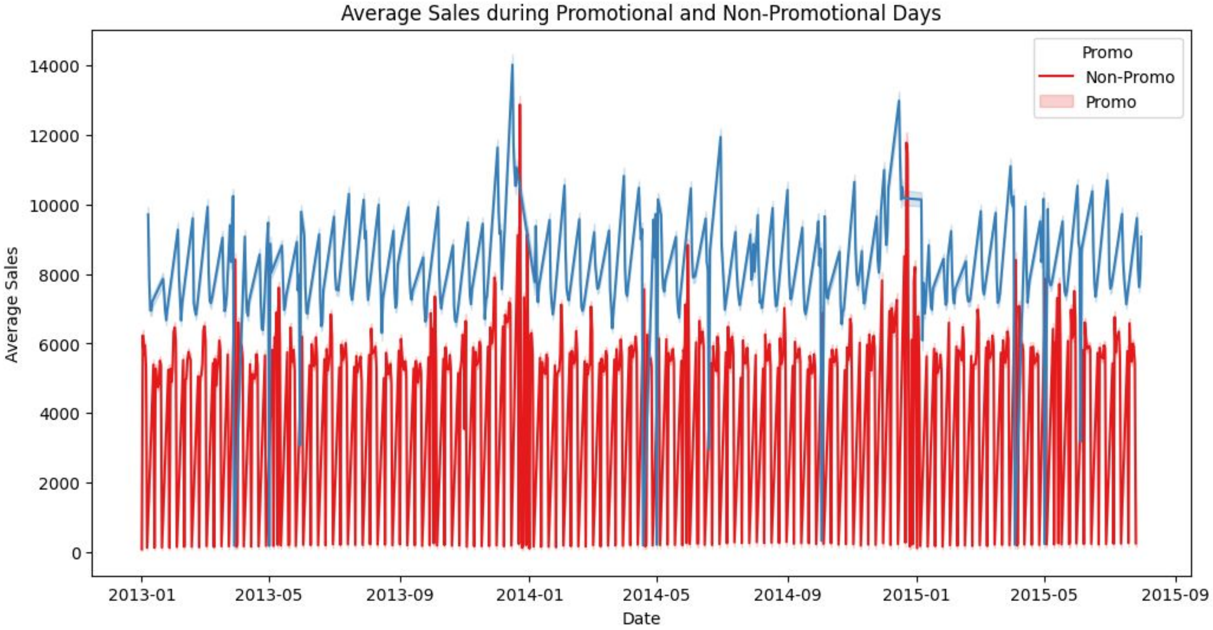
The chart visually represents the central tendency, spread, and skewness of sales data, showcasing the typical sales level, variability across values, and the presence of outliers or extremes in either direction.

NUMBER OF STORES IN EACH STORE TYPE



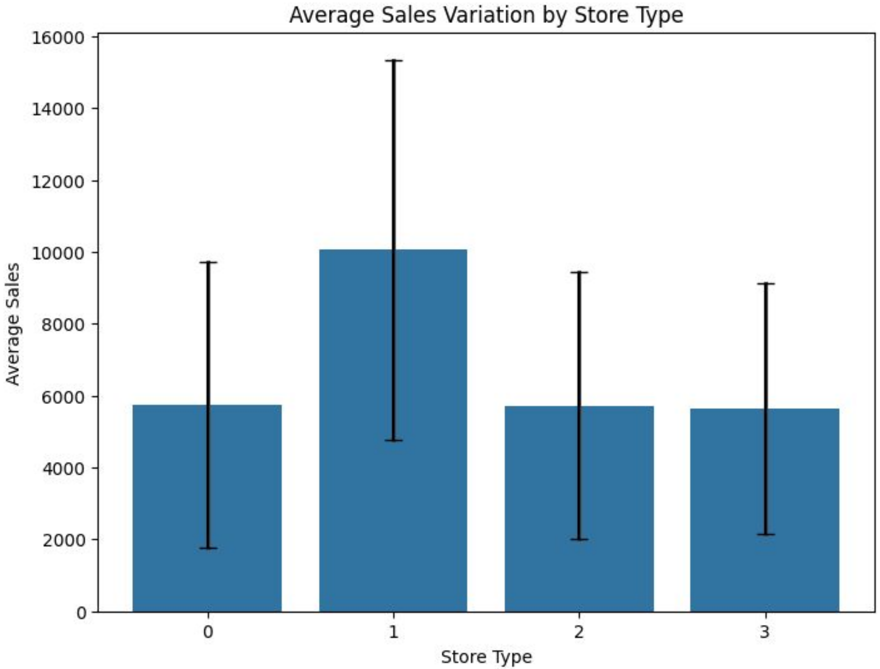
Number of stores are highest in type 0 store and the number of stores which are lowest in type 1 store.

Average Sales during Promotional and Non-Promotional Days



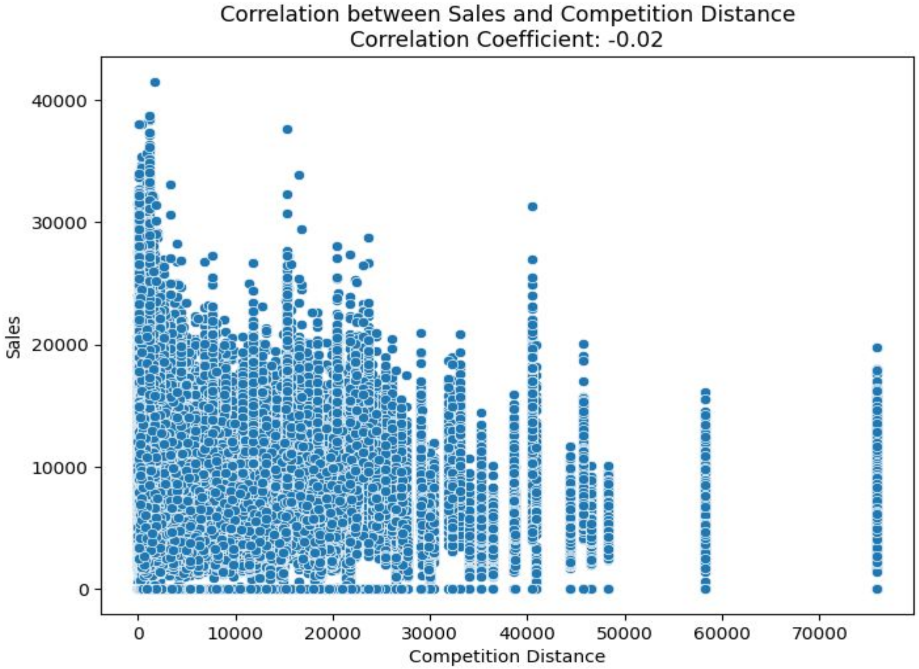
Average sales were highest for both promotional and non promotional in period of 2013-09 to 2014-01.

Average Sales Variation by Store Type



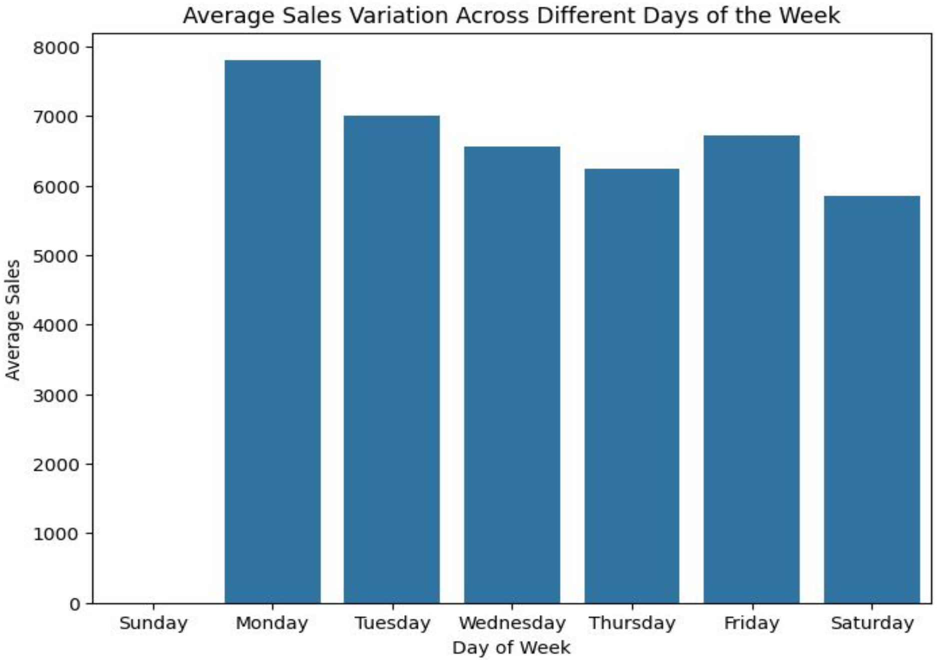
Average sales by store type 1 is highest among all of the other store type.

Correlation between Sales and Competition Distance



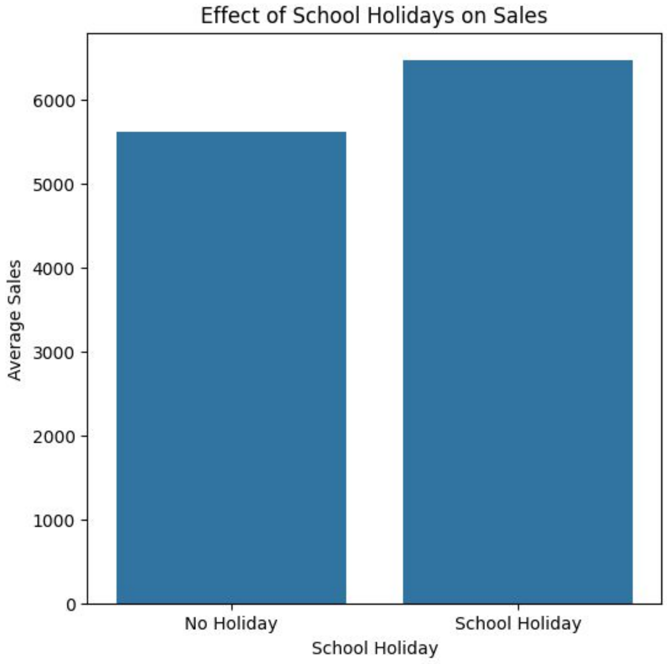
A correlation coefficient of -0.0189 indicates a weak or negligible linear relationship between sales and competition distance suggests limited predictive power, urging consideration of additional factors for effective business decisions in retail.

Average Sales Variation Across Different Days of the Week



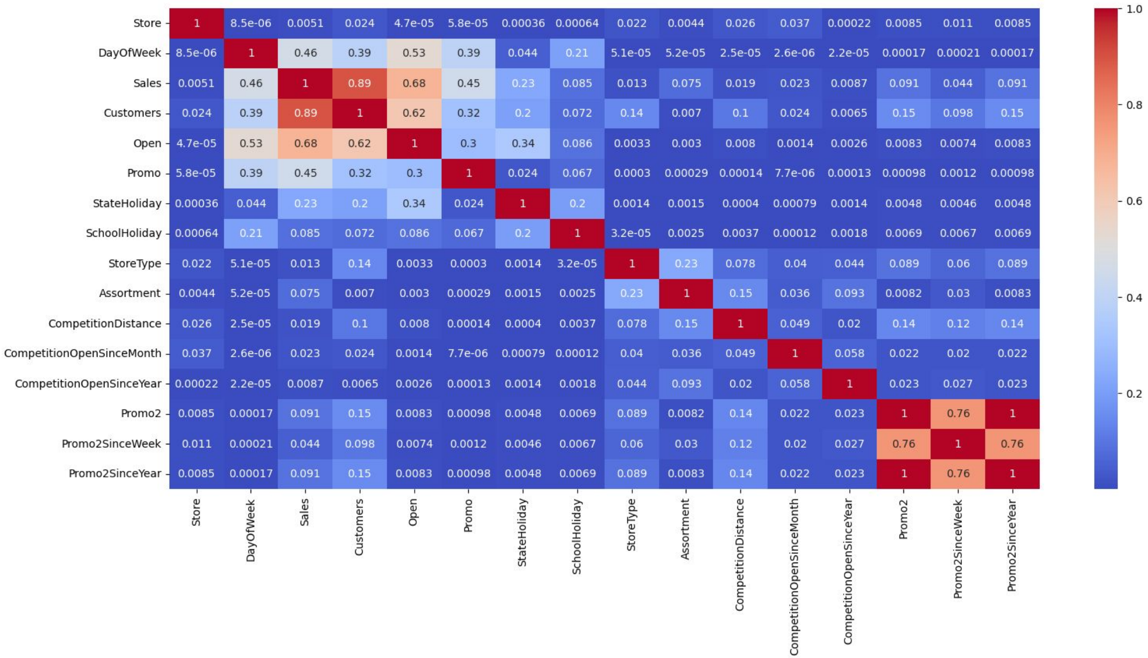
When there is Sunday sales are very low and on Monday sales are highest.

Effect of School Holidays on Sales



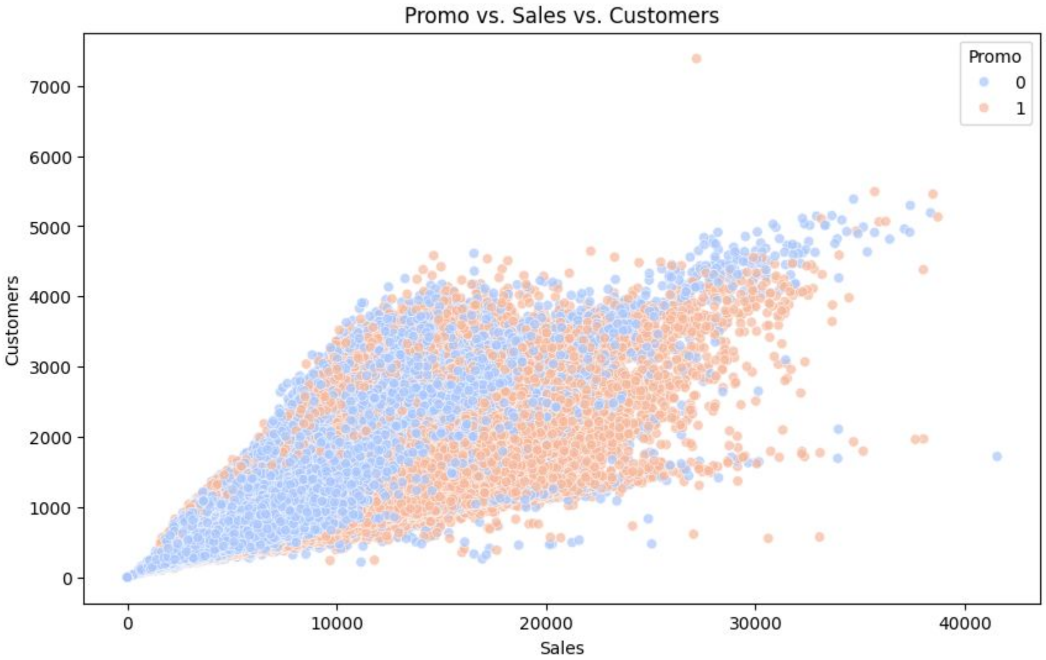
Sales are increased in school holiday in comparison to no Holiday.

CORRELATION MATRIX



Most of the features have negative correlation between them as showed by heatmap.

Promo vs Sales vs Customers



When promo is 1, sales are increasing as there are more customers.

MACHINE LEARNING REGRESSION MODELS IMPLEMENTED

Random Forest Regressor

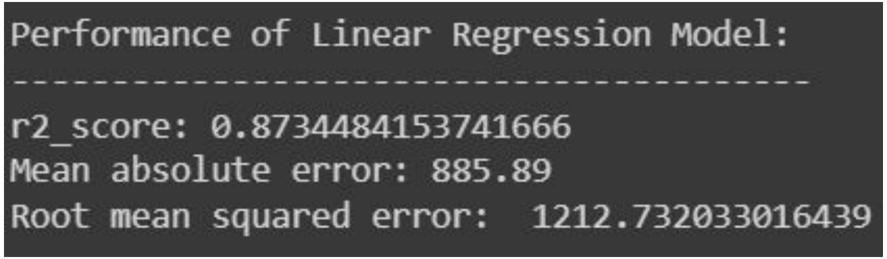
XGBoost

Linear Regression

Decision Trees

REGRESSION MODEL - LINEAR REGRESSION

Linear Regression is a statistical method used to model the relationship between one or more independent variables and a dependent variable by fitting a linear equation to observed data.

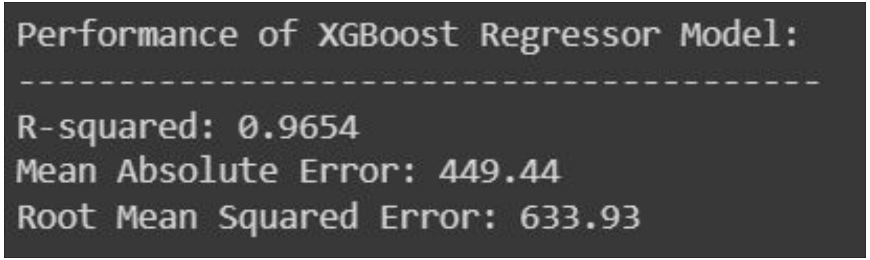


```
Performance of Linear Regression Model:
----------------------------------------
r2_score: 0.8734484153741666
Mean absolute error: 885.89
Root mean squared error:  1212.732033016439
```

Linear Regression

REGRESSION MODEL - XGBOOST

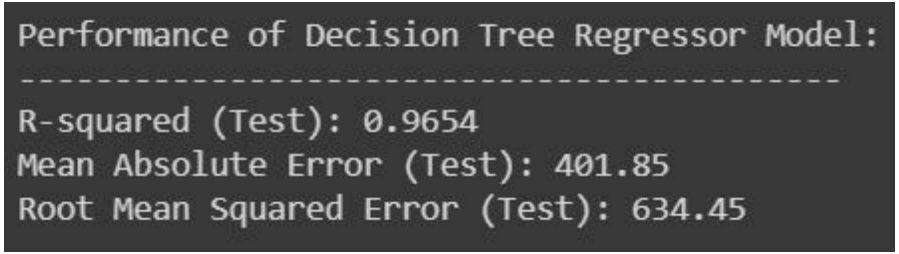
XGBoost builds a series of decision trees sequentially, where each new tree corrects errors made by the previous ones.



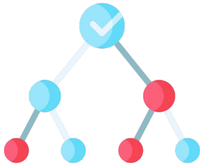
```
Performance of XGBoost Regressor Model:
----------------------------------------
R-squared: 0.9654
Mean Absolute Error: 449.44
Root Mean Squared Error: 633.93
```

REGRESSION MODEL - DECISION TREE

Decision trees predicts the target variable by partitioning the data into regions and assigning a constant value to each region.



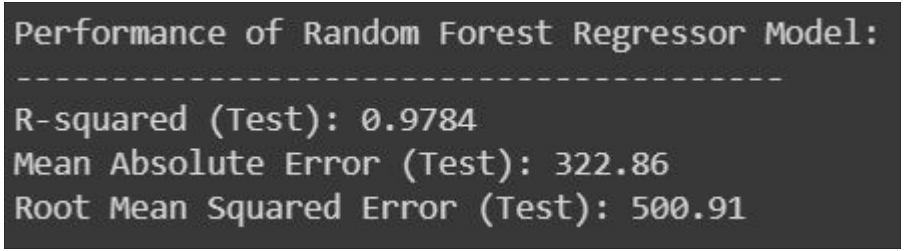
```
Performance of Decision Tree Regressor Model:
---------------------------------------------
R-squared (Test): 0.9654
Mean Absolute Error (Test): 401.85
Root Mean Squared Error (Test): 634.45
```



Decision Trees

REGRESSION MODEL - RANDOM FOREST

Random Forest consists of a collection of decision trees, where each tree is built independently and trained on a random subset of the training data.



```
Performance of Random Forest Regressor Model:
----------------------------------------
R-squared (Test): 0.9784
Mean Absolute Error (Test): 322.86
Root Mean Squared Error (Test): 500.91
```

Random Forest Regressor

MODEL SELECTION

Model Selected :

Random Forest Regressor

Best Hyperparameter :
R-squared (Test): 0.9795
Mean Absolute Error (Test): 314.08
Root Mean Squared Error (Test): 488.38

Random Forest Regressor

Random Forest Regressor is selected for the regression problem by considering the evaluation metrics.

CONCLUSION

There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.

The positive effect of promotion on Customers and Sales is observable.

Most stores have competition distance within the range of 0 to 10 kms and had more sales than stores far away probably indicating competition in busy locations vs remote locations.

The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.

Out of the four methods, Random Forest proved to be the most accurate, achieving a R2_Score of 0.9795, MAE of 314.08 and RMSE of 488.38.

Thank You