Capstone Project - 2

Retail Sales Prediction
(Supervised ML Regression)

Team Members
Nihal Habeeb
Parvez Makandar

Content

- Problem Description
- Data Summary
- Exploratory Data Analysis
- Fitting the Models
- Model Selection
- Feature Importance
- Conclusion

Problem Description

- Rossmann has over 3000 drug stores in 7 European countries.
- The managers are tasked with predicting their sales for six weeks in advance.
- We have historical sales data for 1,115 Rossman stores.
- There are information on many parameters and the task is to forecast or predict the sales based on these parameters.

Problem Description

Predicting the daily sales is a necessary step for a business in order to estimate the cost and revenue as well as predict the rise and fall of demands well in advance. It might also help in early identification of business problems, such as effects of marketing methods, effect of neighboring competitors etc.

Data Summary

- Customer : - The Number of customers on a given day in a store.
- State Holiday :- Indicates a state holiday.
- Store Type : Differentiate between 4 different store models.
- Assortment : Describes an assortment level i.e a : basic, b : extra and c : extended.
- Competition Distance : Distance in meters to the nearest competition store.

Data Summary (Contd.)

7. CompetitionOpenSince[Year/Month] :- Gives the approximate year and month of the time the nearest competitor is opened.

8. Promo :- Indicates whether a store is running a promo on that day.

9. Promo2 :- Indicates whether a store is continuing promotion.

10. Promo2Since[Year/Week] :- Gives the approximate year and calender week of the time when the store started participating in Promo2.

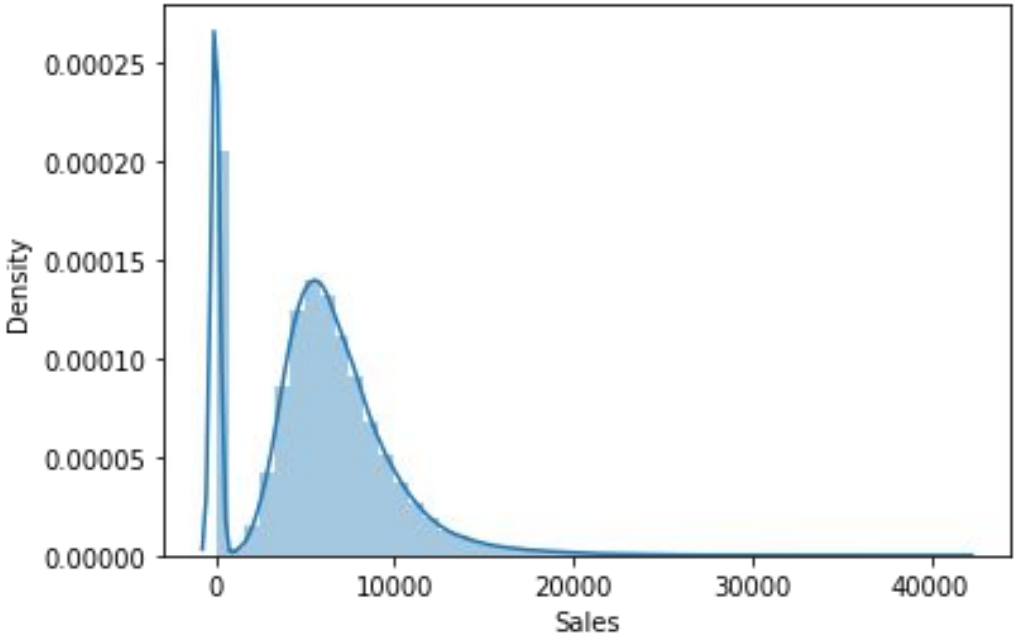
11. PromoInterval :- Describes an interval or name of months when the store runs Promo2.

Data Cleaning

- Dropped columns with high null values count: CompetitionOpenSinceMonth, CompetitionOpenSinceYear, Promo2SinceWeek, Promo2SinceYear, PromoInterval
- Replaced null values of CompetitionDistance with mean value.
- Joined the two data frames

EDA

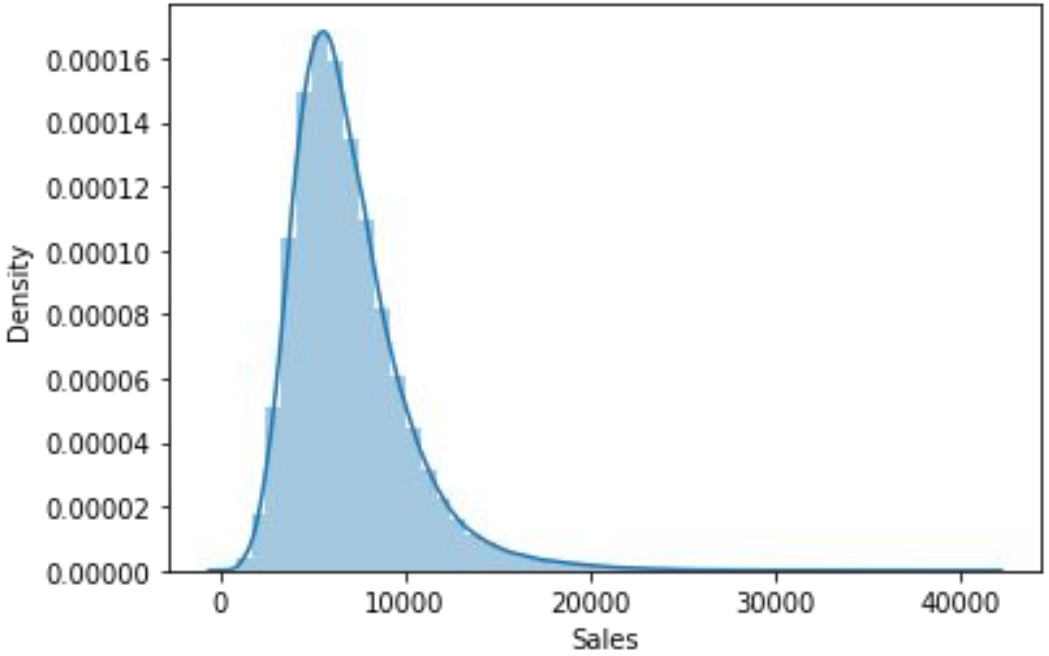
Sales variable distribution



- A lot of values at 0
- Data includes closed stores

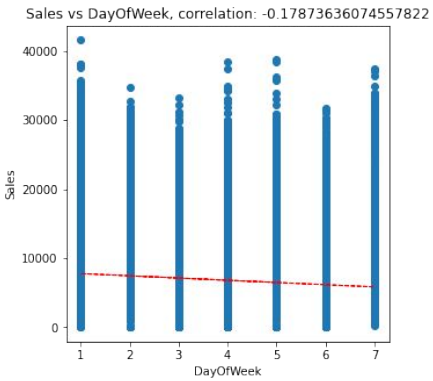
EDA

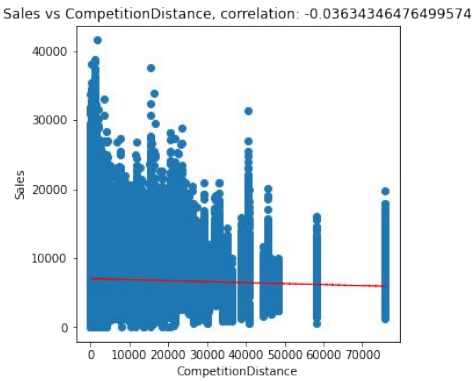
Sales variable distribution

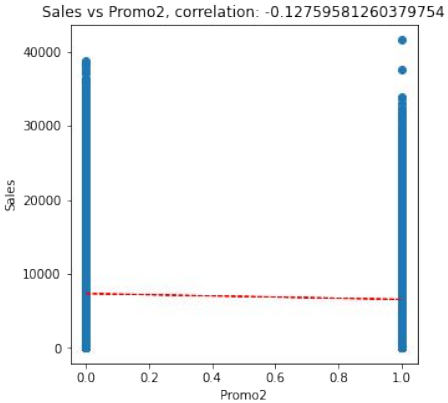


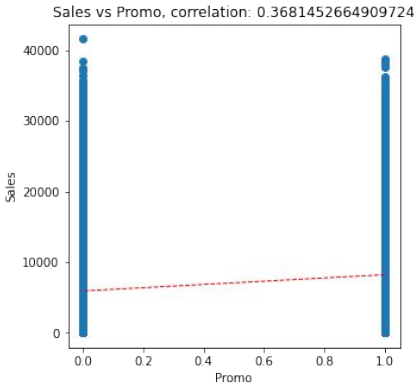
- New dataset only considering Open stores
- Peak at 0 is removed

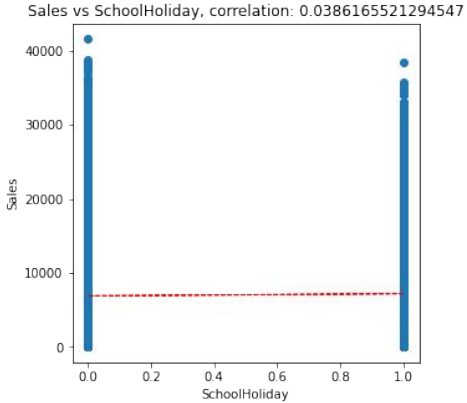
Linear Relationship



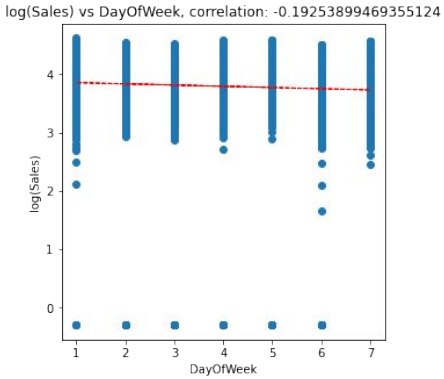


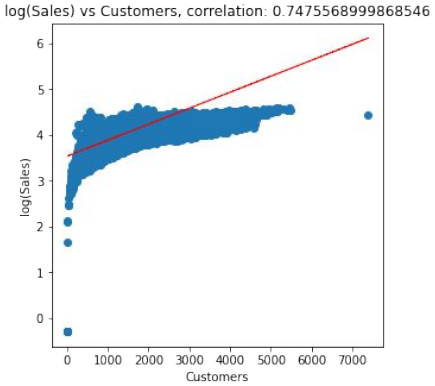


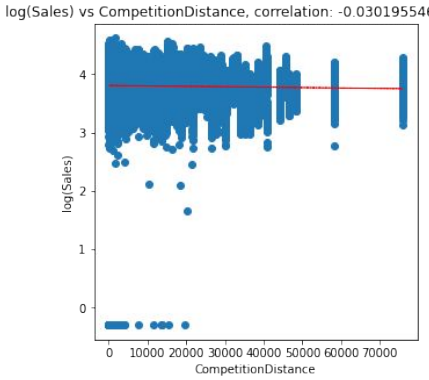


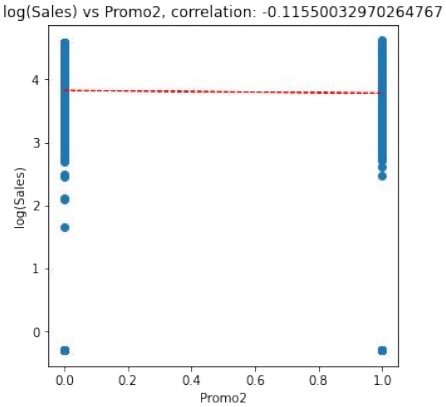


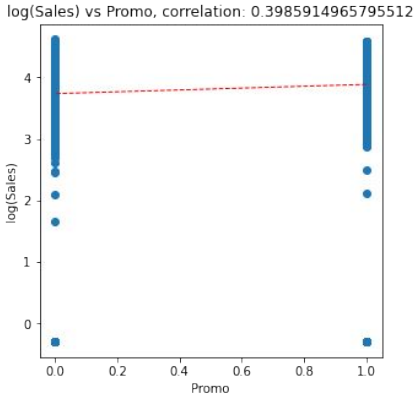
Linear Relationship

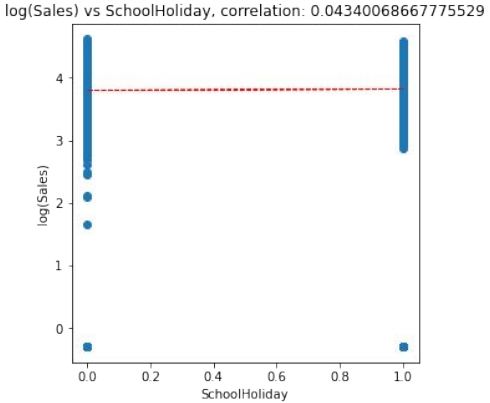




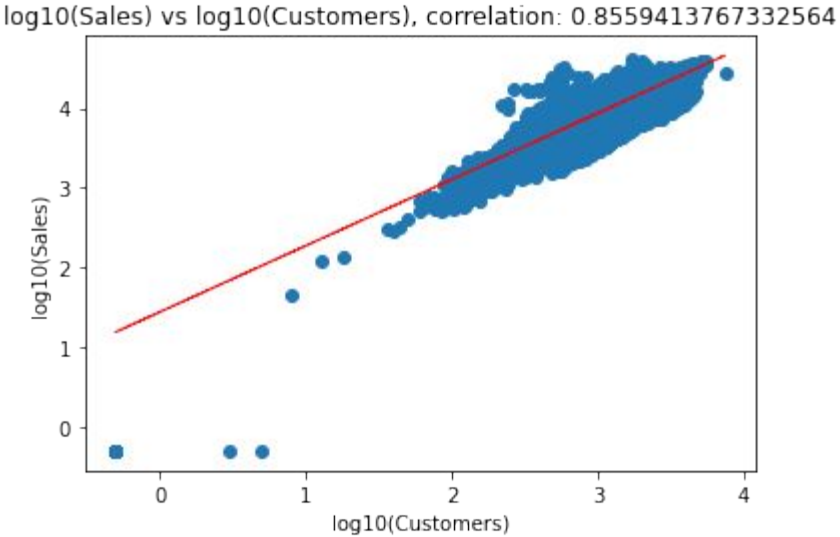






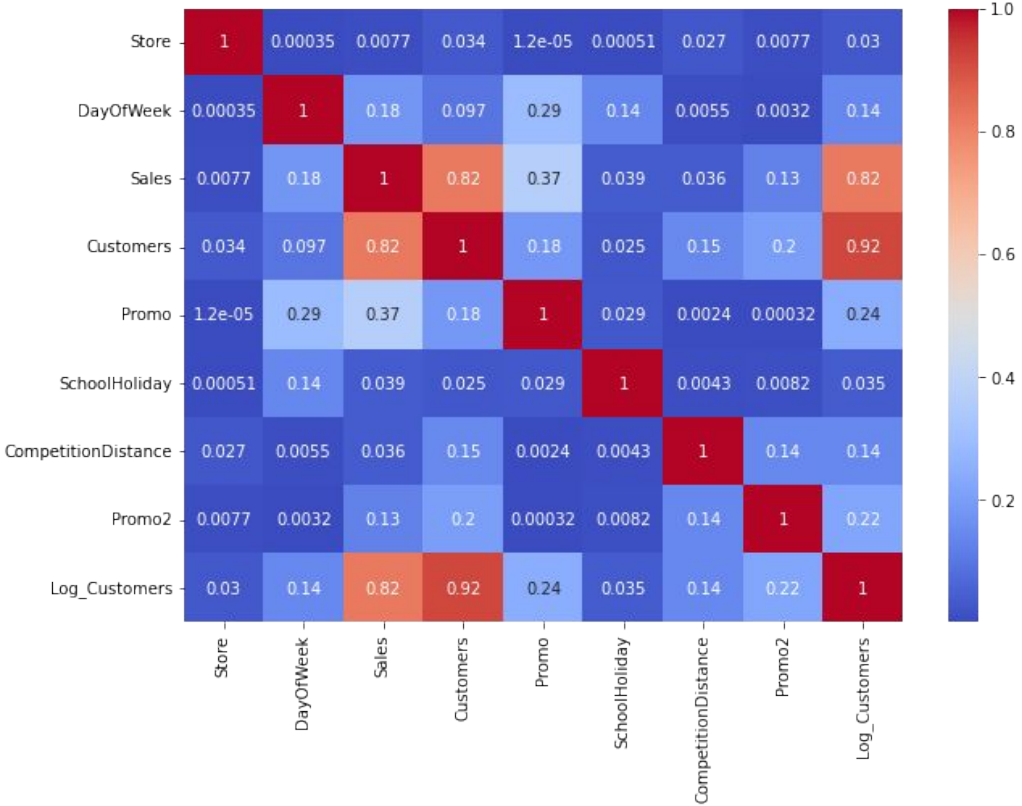


Linear Relationship



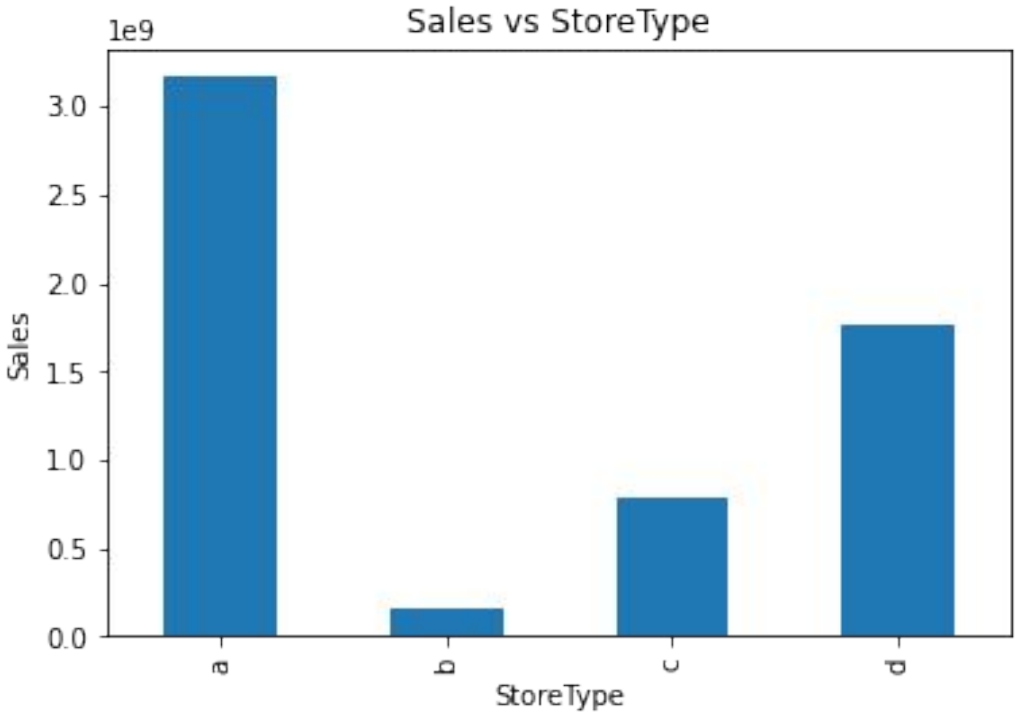
Log transformation of Sales and Customers variable is taken.

Correlation



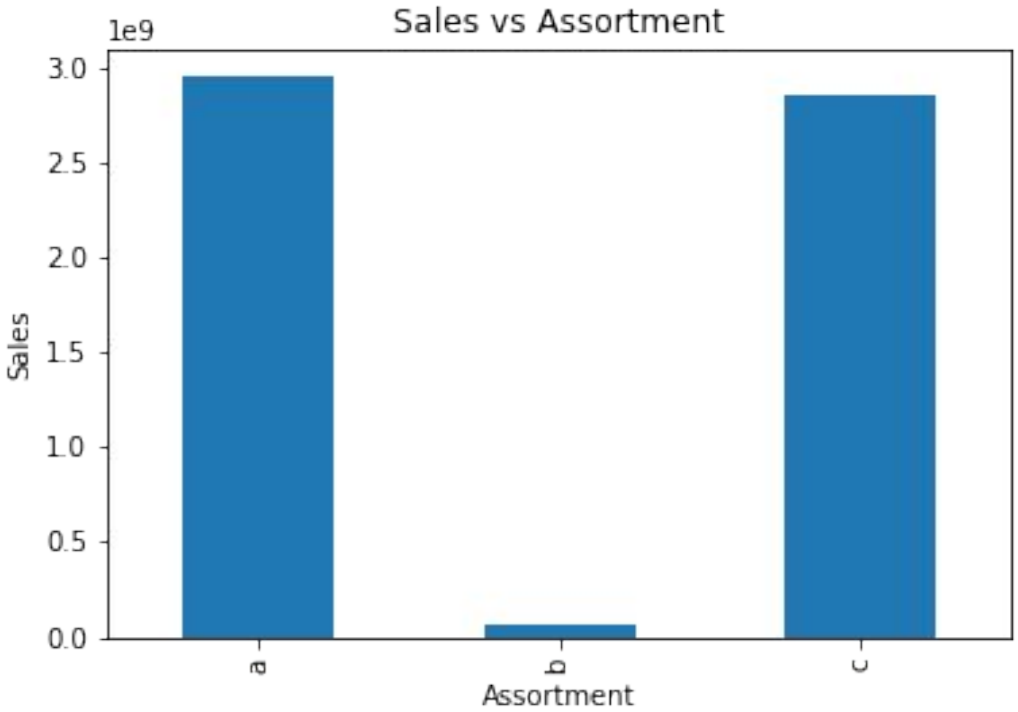
- The independent variables do not show high multicollinearity.
- Target (sales) variable and customers variable show good correlation.

Sales Across Store Type



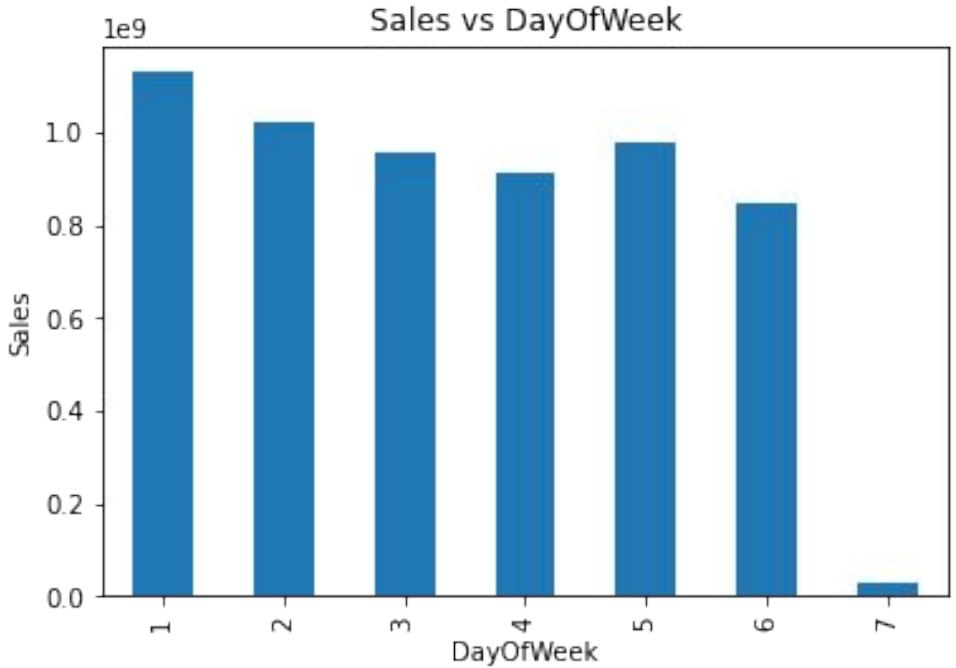
Store type a followed by d has the most sales

Sales Across Assortment



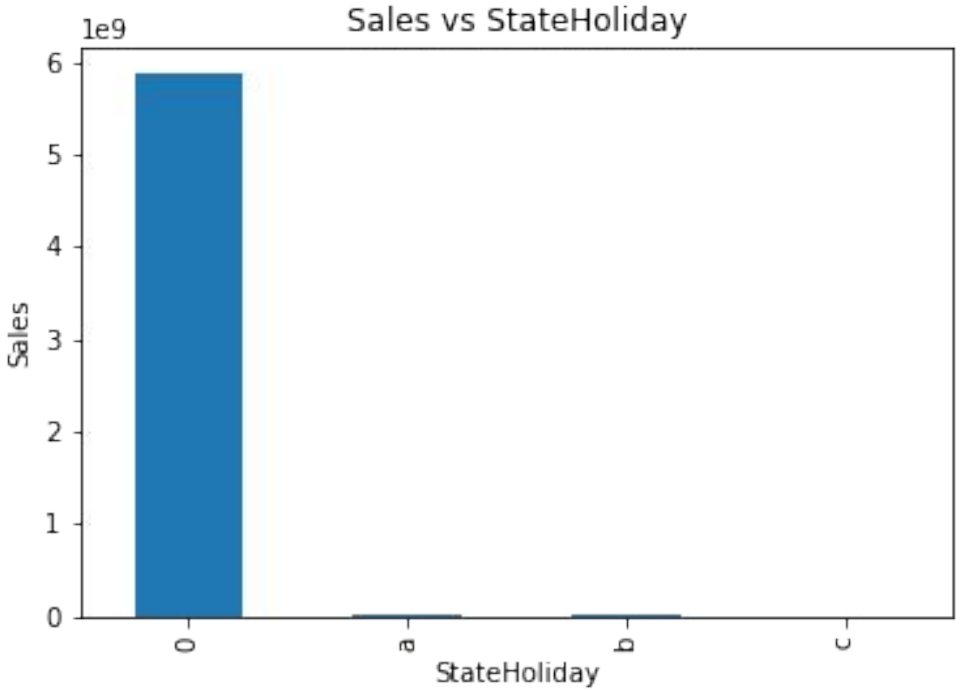
Assortment a and c has fairly high sales

Weekly Sales Trend



Sunday has very low sales compared to the rest of the days of the week.

Sales Across Holidays



Most of the sales are not on holidays

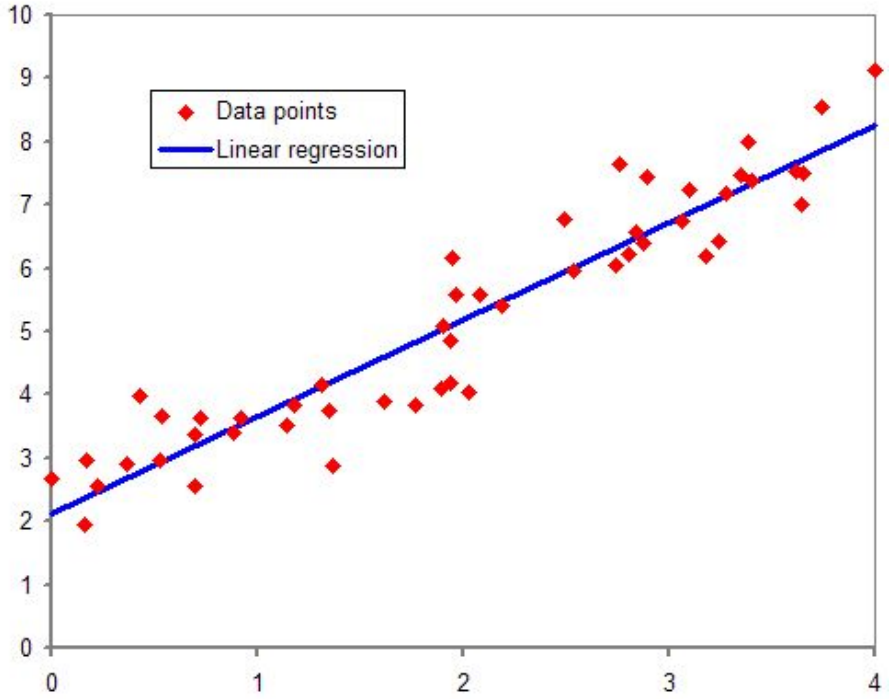
Encoding & Feature Selection

- The binary categorical variables (such as Promo, Promo2, SchoolHoliday) are already in numerical form.
- DayOfWeek variable is also in numerical form (1 to 7 for all the days in a week).
- One Hot encoding is done on the other categorical variables (StateHoliday, StoreType, Assortment).
- From date variable, month information is retrieved and only that is used in the models.
- log(Customers) is used instead of Customers variable.

Models Used For Prediction

- Linear Regression
  - Not regularized (OLS) model
  - Lasso
  - Ridge
  - Elastic Net
- Decision Tree Regression
- Random Forest Regression

Linear Regression

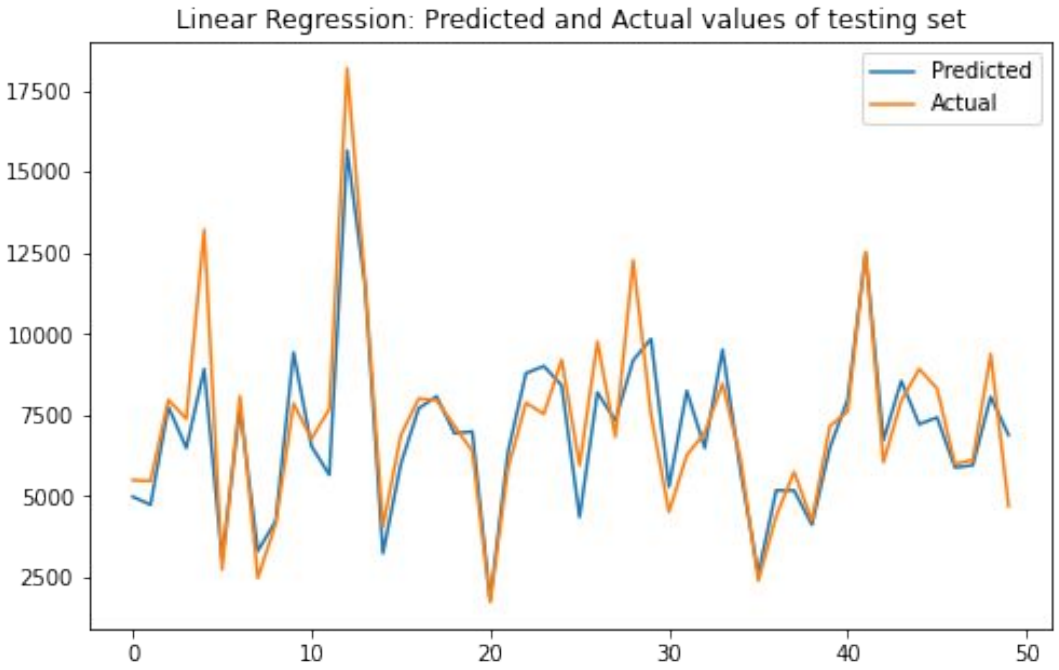


www.wikipedia.org

$$y = \beta_0+\beta_1x_1+\beta_2x_2+..+\beta_px_p$$

- Linear relationship between independent and dependent variables is assumed.
- Ordinary Least squares method is used to find the best fit line.

Linear Regression



Model Performance on Training Data

Root Mean Squared Error: 1213.89

R2 Score: 0.847

Adjusted R2 Score: 0.847

Model Performance on Testing Data

Root Mean Squared Error: 1212.66

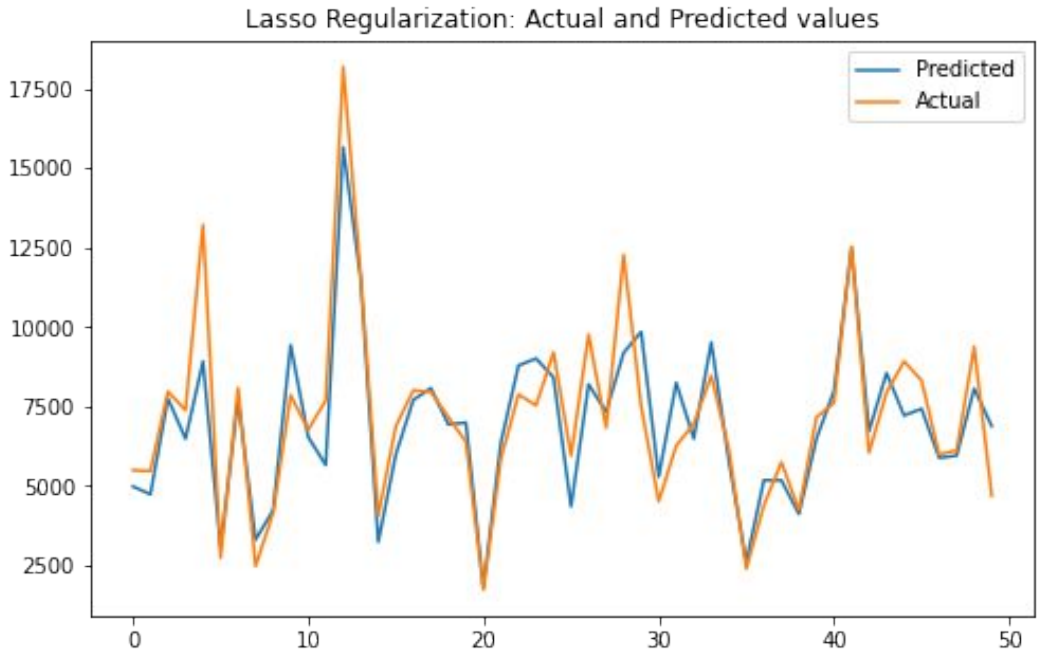
R2 Score: 0.846

Adjusted R2 Score: 0.846

Regularization

- Ridge regularization $RSS + \lambda\sum\beta_j^2$
- Lasso regularization $RSS + \lambda\sum |\beta_j|$
- Elastic Net regularization $RSS + \lambda\sum |\beta_j| + \lambda\sum\beta_j^2$

Lasso Regression



Model Performance on Training Data

Root Mean Squared Error: 1213.89

R2 Score: 0.847

Adjusted R2 Score: 0.847

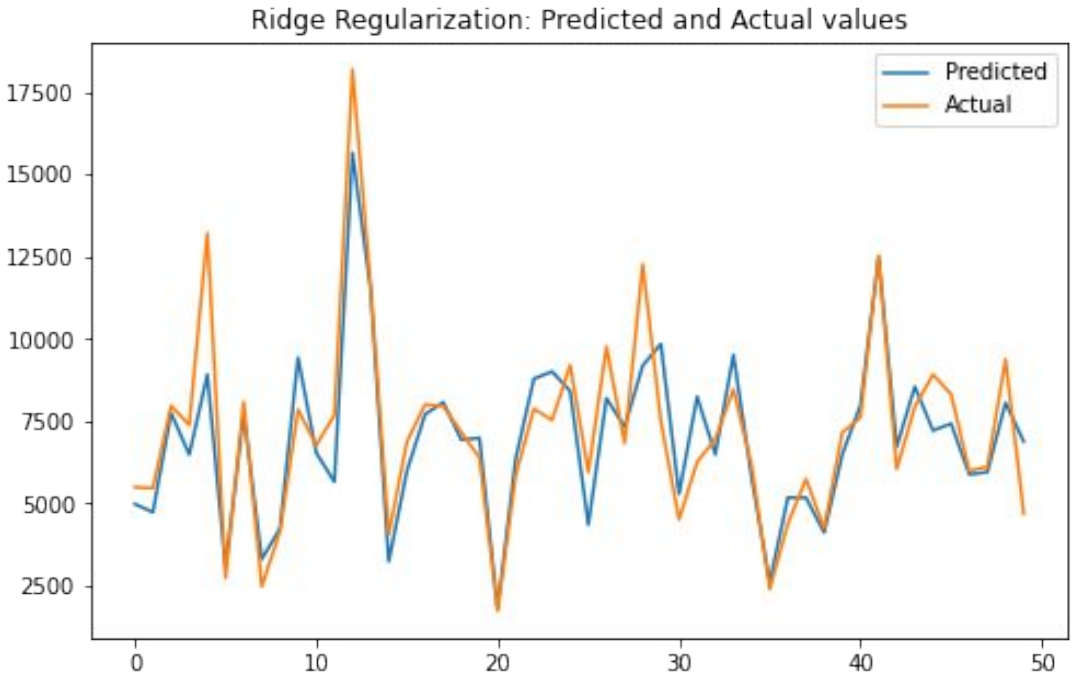
Model Performance on Testing Data

Root Mean Squared Error: 1212.65

R2 Score: 0.846

Adjusted R2 Score: 0.846

Ridge Regression



Model Performance on Training Data

Root Mean Squared Error: 1213.89

R2 Score: 0.847

Adjusted R2 Score: 0.847

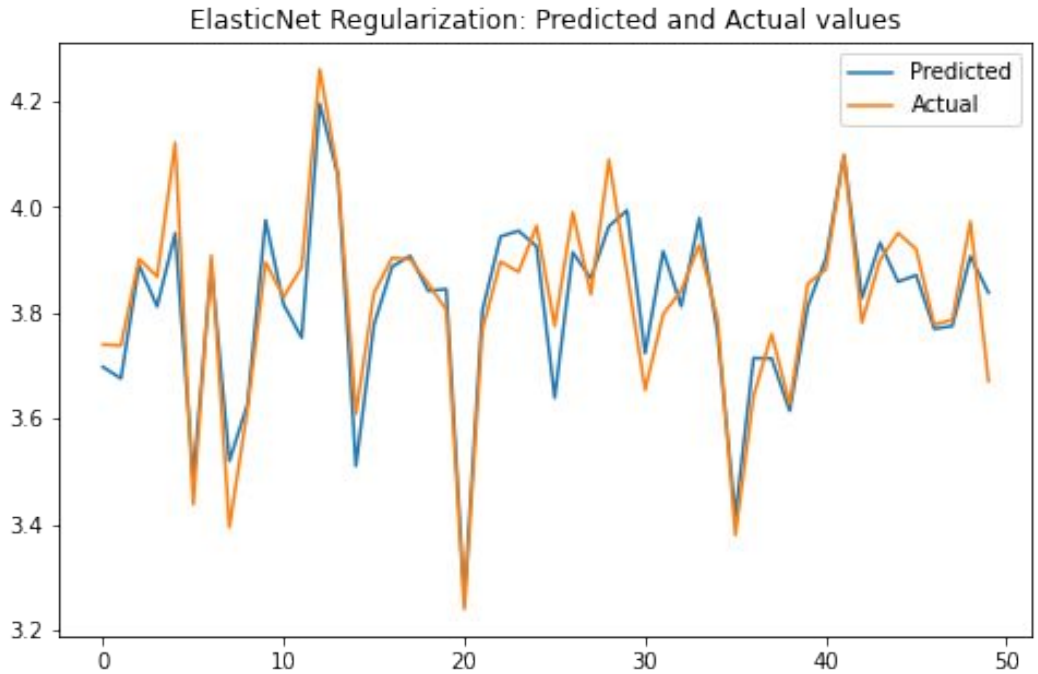
Model Performance on Testing Data

Root Mean Squared Error: 1212.65

R2 Score: 0.846

Adjusted R2 Score: 0.846

Elastic Net Regression



Model Performance on Training Data

Root Mean Squared Error: 1213.89

R2 Score: 0.847

Adjusted R2 Score: 0.847

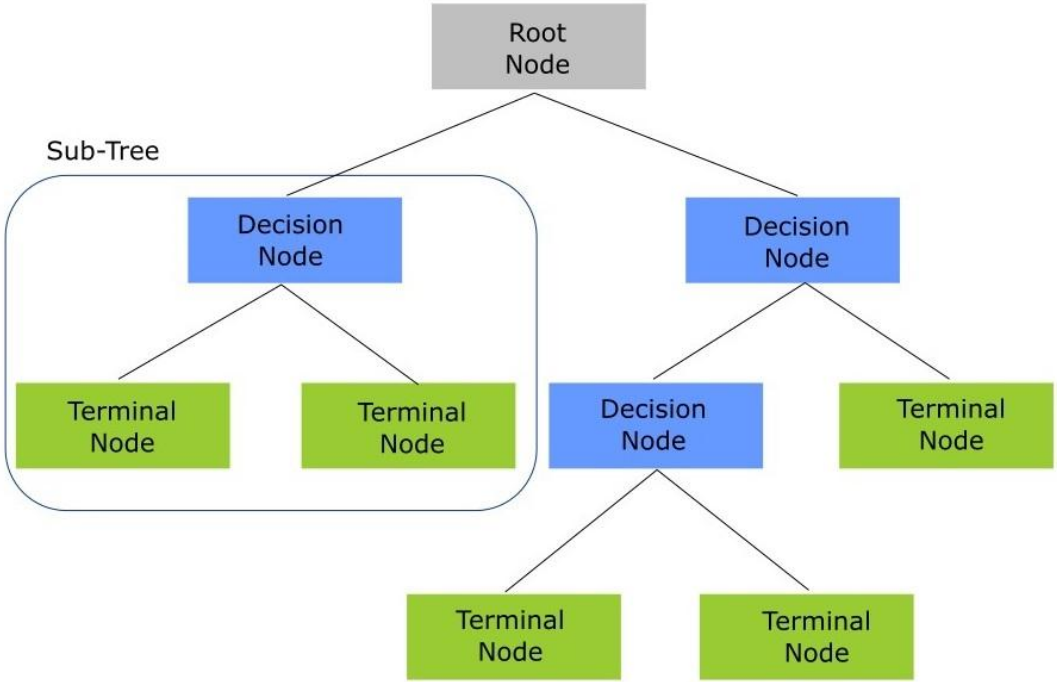
Model Performance on Testing Data

Root Mean Squared Error: 1212.65

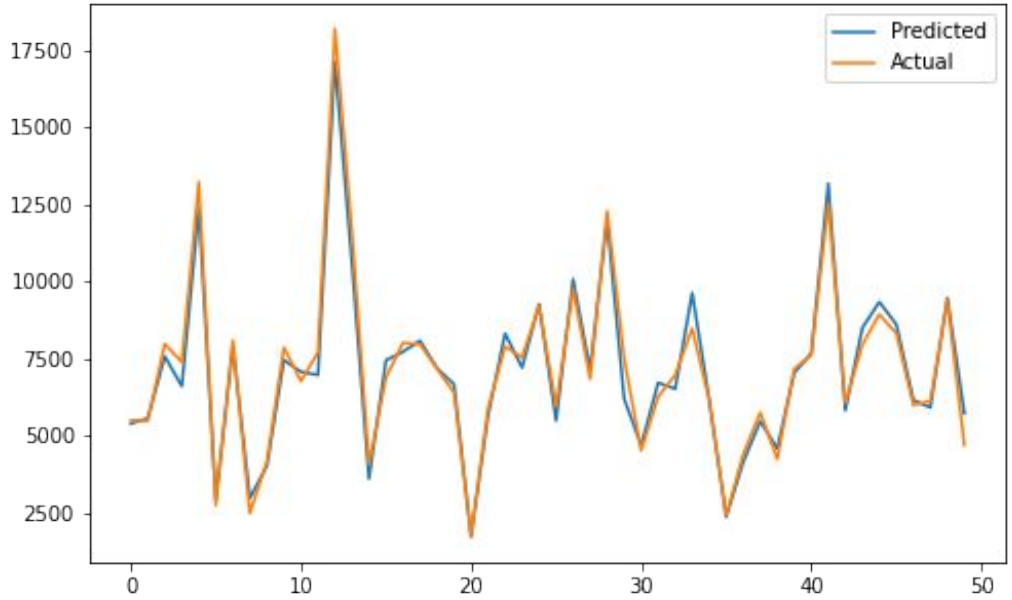
R2 Score: 0.846

Adjusted R2 Score: 0.846

Decision Tree Regression



Decision Tree Regression



Model Performance on Training Data

Root Mean Squared Error: 570.47

R2 Score: 0.966

Adjusted R2 Score: 0.966

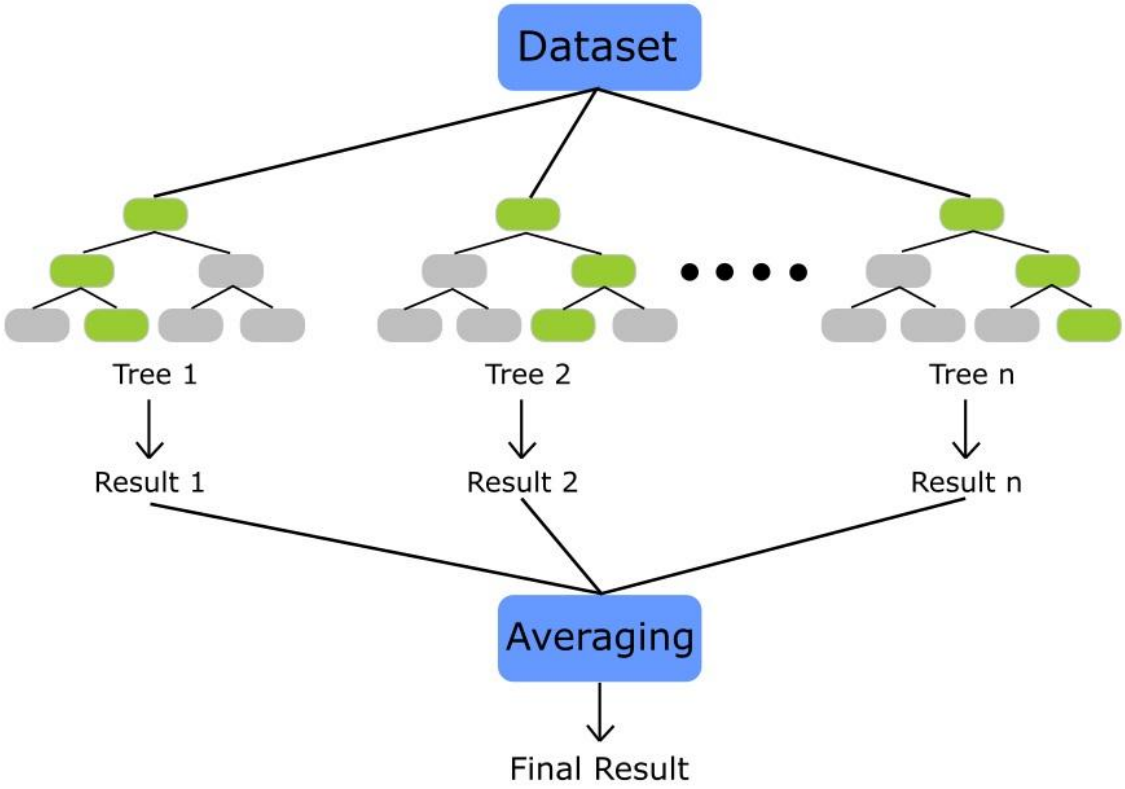
Model Performance on Testing Data

Root Mean Squared Error: 680.98

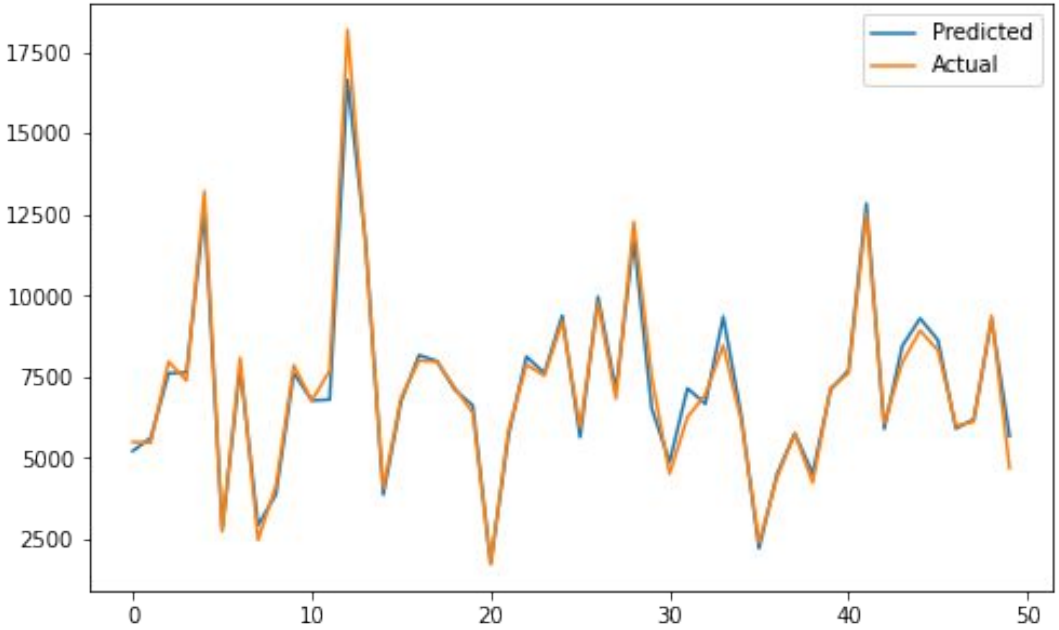
R2 Score: 0.951

Adjusted R2 Score: 0.951

Random Forest Regression



Random Forest Regression



Model Performance on Training Data

Root Mean Squared Error: 475.41

R2 Score: 0.976

Adjusted R2 Score: 0.976

Model Performance on Testing Data

Root Mean Squared Error: 599.68

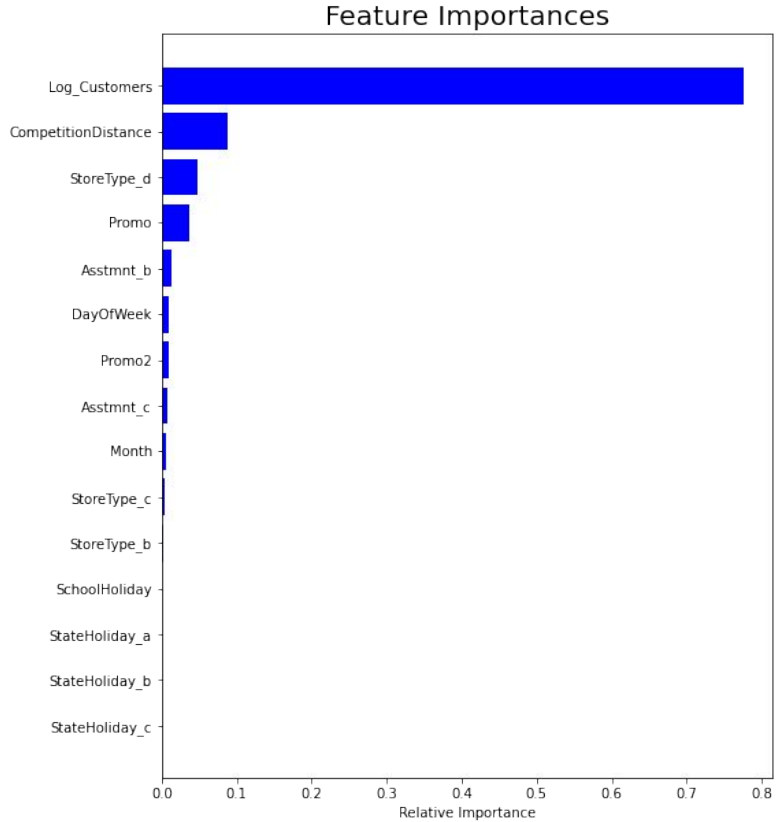
R2 Score: 0.962

Adjusted R2 Score: 0.962

Evaluation of Models

Models	R2 Score	RMSE
Linear Regression	0.846	1212.659
Lasso Regression	0.846	1212.659
Ridge Regression	0.846	1212.659
Elastic Net Regression	0.846	1212.659
Decision Tree Regression	0.951	680.899
Random Forest Regression	0.962	599.376

Feature Importances



Customers followed by competition distance were most important in the construction of Random Forest model.

Conclusion

- Random Forest regression was found to be the best performing model with an R2 score of 0.962
- Decision tree regression comes close second with an R2 score of 0.951
- All the linear regression models have a similar performance with an R2 score of 0.846
- The regularized linear models did not have an improved performance

Thank You!!