Capstone Project

Retail Sales Prediction

Individual Project

Name : Aakash Ramnani

Content

- Problem Statement
- Data Summary
- Project Summary
- EDA (Exploratory Data Analysis)
- Hypothesis Testing
- Feature Engineering and Data Pre-processing
- Model Implementation
  1. Linear Regression - OLS, Lasso, Ridge, Elasticnet
  2. Decision Trees
  3. Random Forest
- Hyperparameter Tuning
- Model Performance and Evaluation
- Conclusion

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. 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.

Data Summary

- Id - an Id that represents a (Store, Date) duple within the set
- Store - a unique Id for each store
- Sales - the turnover for any given day (Dependent Variable)
- 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. An assortment strategy in retailing involves the number and type of products that stores display for purchase by consumers.
- 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

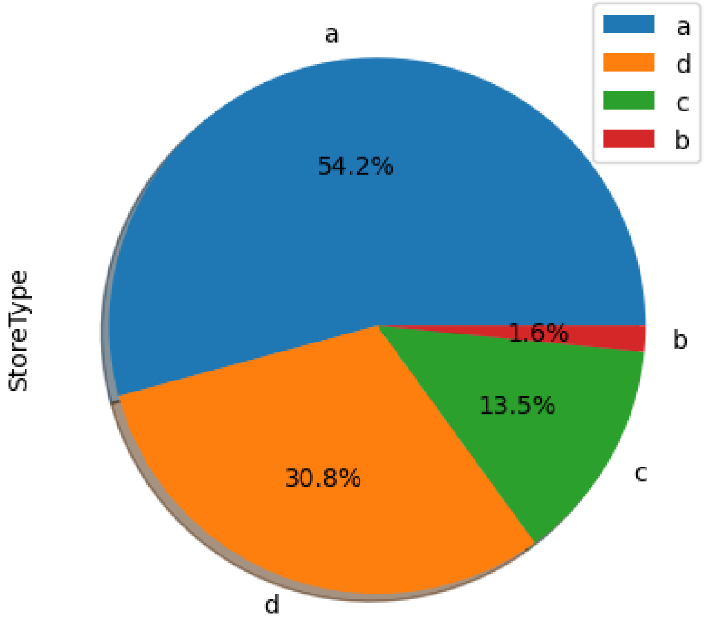
Project Summary

- The goal is to predict the Sales of a given store on a given day.
- The main steps of the project are:
- **Basic EDA(Exploratory Data Analysis):** I have performed basic EDA to understand the data and its characteristics. also, have used Univariate - BI variate - and Multivarite analysis to Understanding the correlation between variables and to explore distribution and intercation of variables.
- **Hypothesis Testing:** I have performed hypothesis testing to test the relationship between the variables. Also have used statistical tests such as t-test, ANOVA, chi-square test. to compare the means or proportions of different groups or categories.
- **Feature Engineering and Data Preprocessing:** To create and select the most relevant and informative features for the model. Have used methods such as correlation analysis and VIF(Variance Inflation Factor) for feature selection.
- **ML Model Development and Evaluation:** In this project Have developed and evaluated different machine learning Models. Have used both Linear and Polynomial Regression. Also have used Tree Based Algorithm.
- **Model Interpretation and Explanation :** Explained the model predictions using feature importance. It helps to analysis the effect of different factors on the sales.
- **Conclusion**

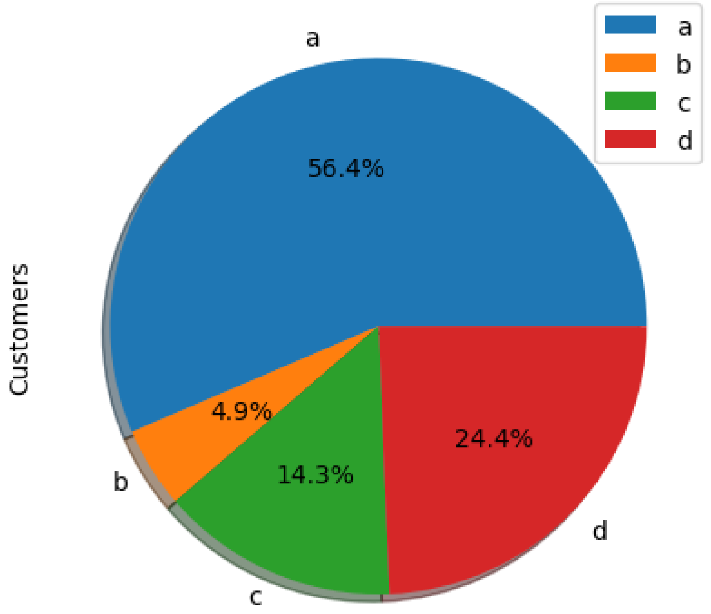
EDA(Exploratory Data Analysis)

- We have divided EDA into 4 Parts
  - Storewise analysis
  - Timewise analysis
  - Analysis with respect to customer
  - Analysis with respect to competition

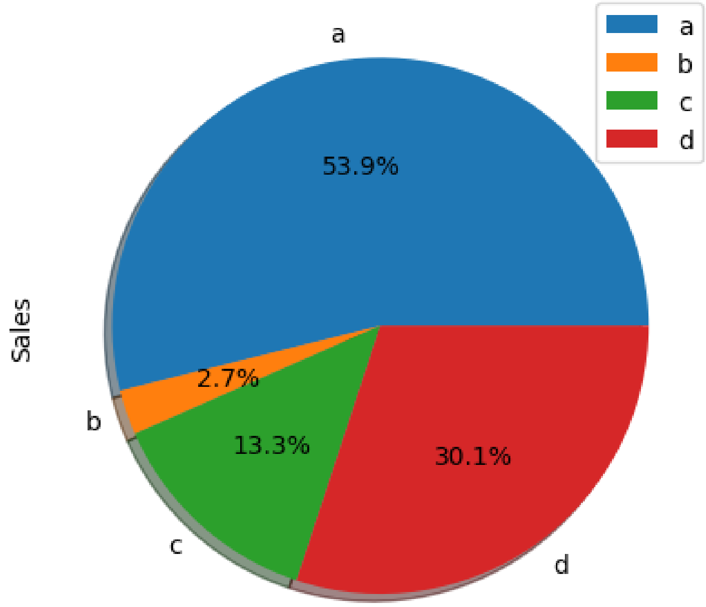
Storewise Analysis



Share of Store Types



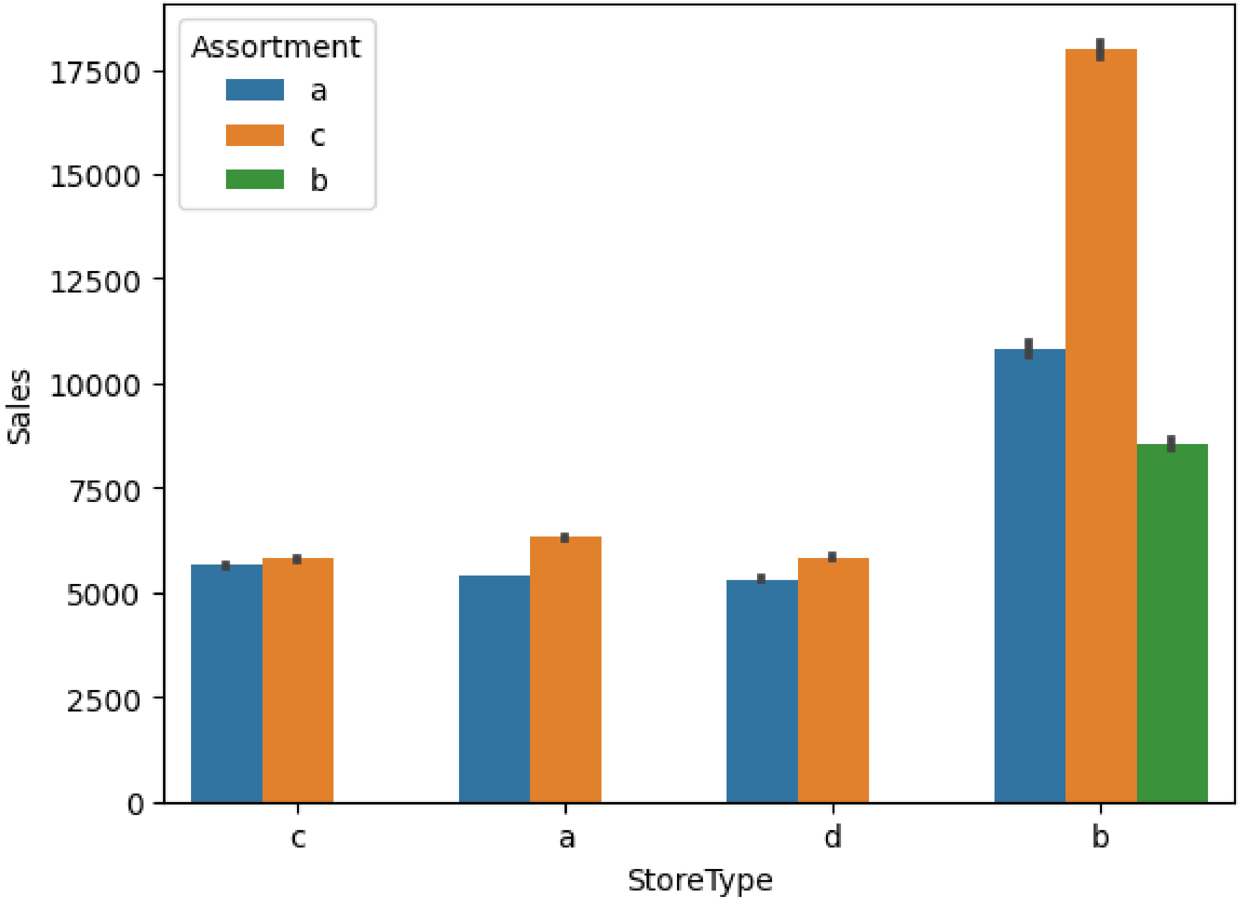
Store type and Customer



Store Type and Sales

- Store type 'a' is more in compare to other store type. Store type 'b' are least in number.
- As discussed earlier Store type 'a' has more customer visits because there are more number of stores with type ‘a’.
- Number of stores with type 'a' are more, probably the reason for more total sales are seen for store type 'a'.

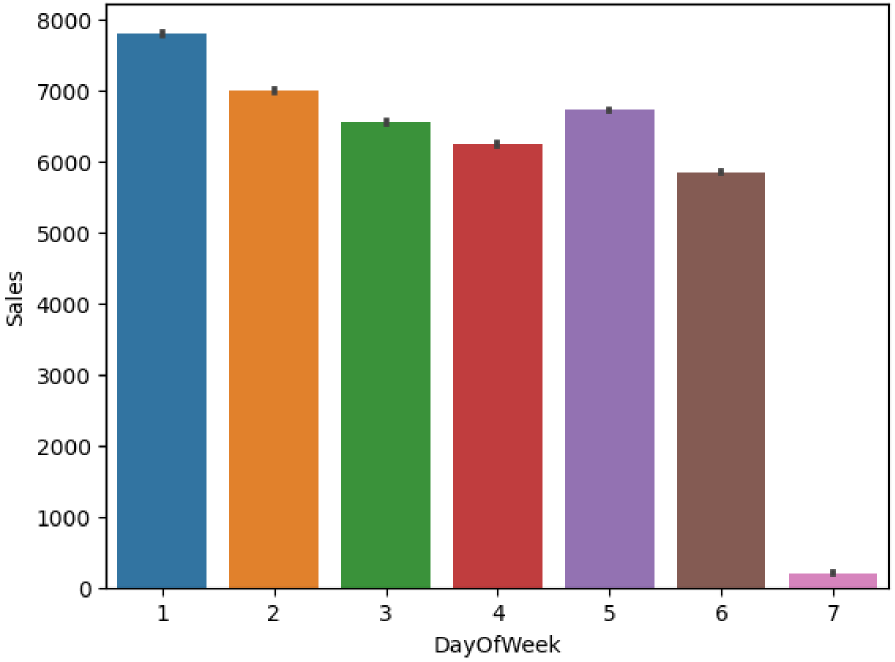
Storewise Analysis (Contd..)

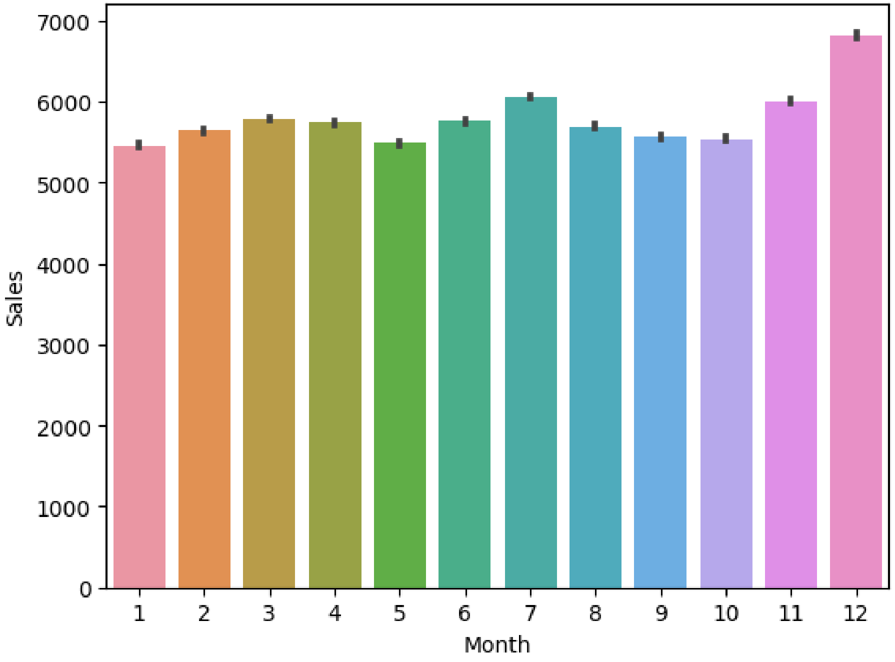


- This bar plot shows that the store types 'a', 'c' and 'd' have only assortment level 'a' and 'c'. On the other hand the store type 'b' has all the three kinds of assortment strategies, a reason why average sales are high for store type 'b' stores.

- Store type 'b' with highest average sales and per store revenue generation looks healthy and a reason for that would be all three kinds of assortment strategies involved.

Timewise Analysis

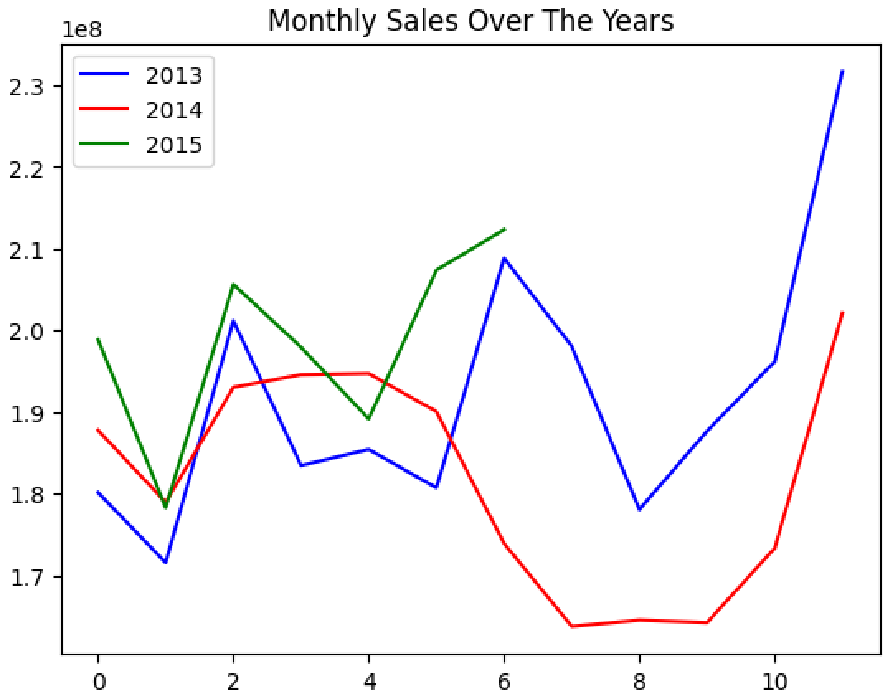


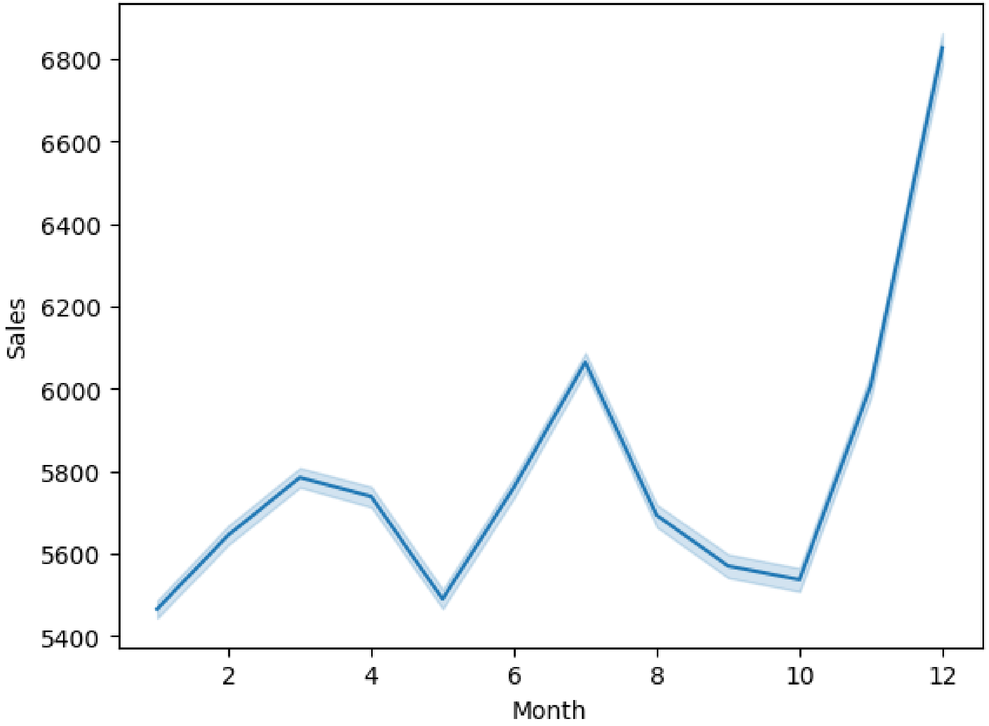


- The graph tells us the store can expect higher sales on Monday, and the average sales are slightly lower on 4th and 6th day of the week and extremely low on Sundays.

- Highest average sales are seen in month of December, July and November.

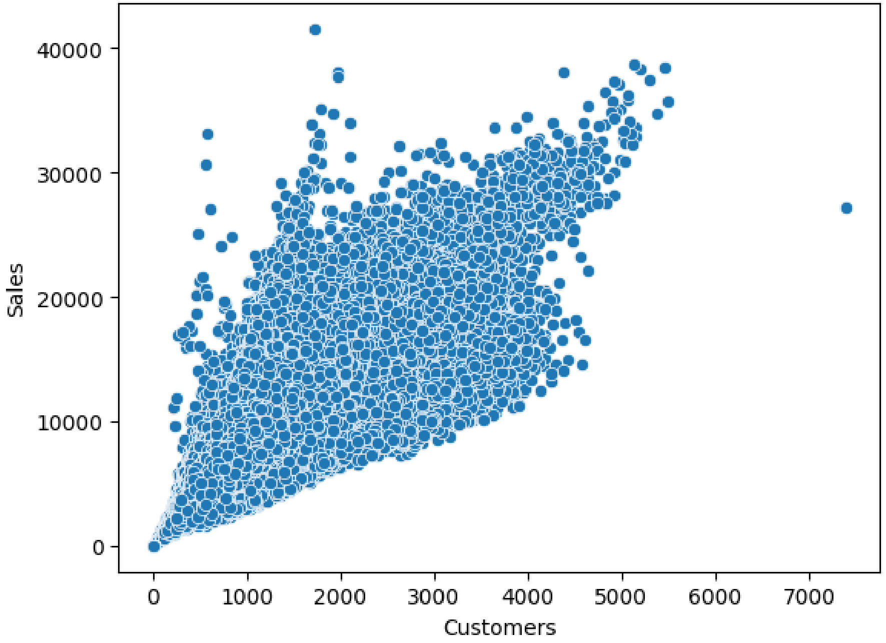
Timewise Analysis(Contd..)

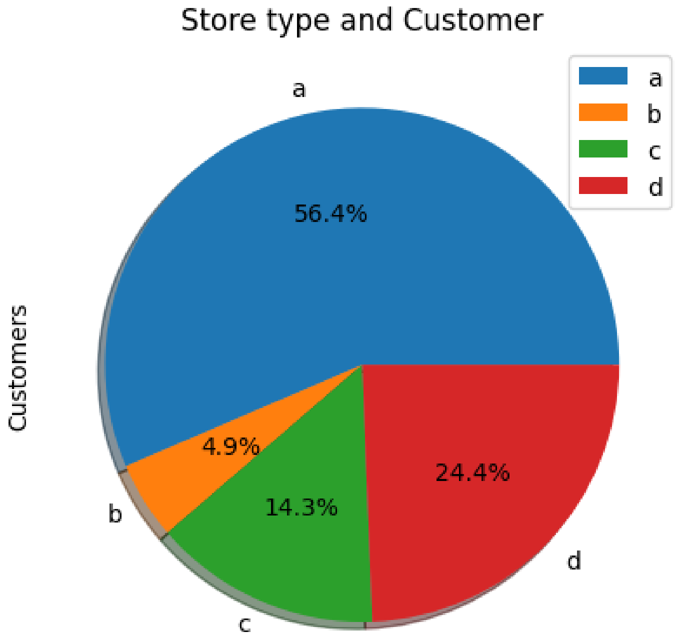




- The Monthly trend for 2013 and 2015 is almost similar. However there is sudden drop in month of July, August and September for year 2014.
- The sudden drop in the month of July, August and September for 2014 is because of the stores were closed for refurbishment which also mentioned in the problem statement.

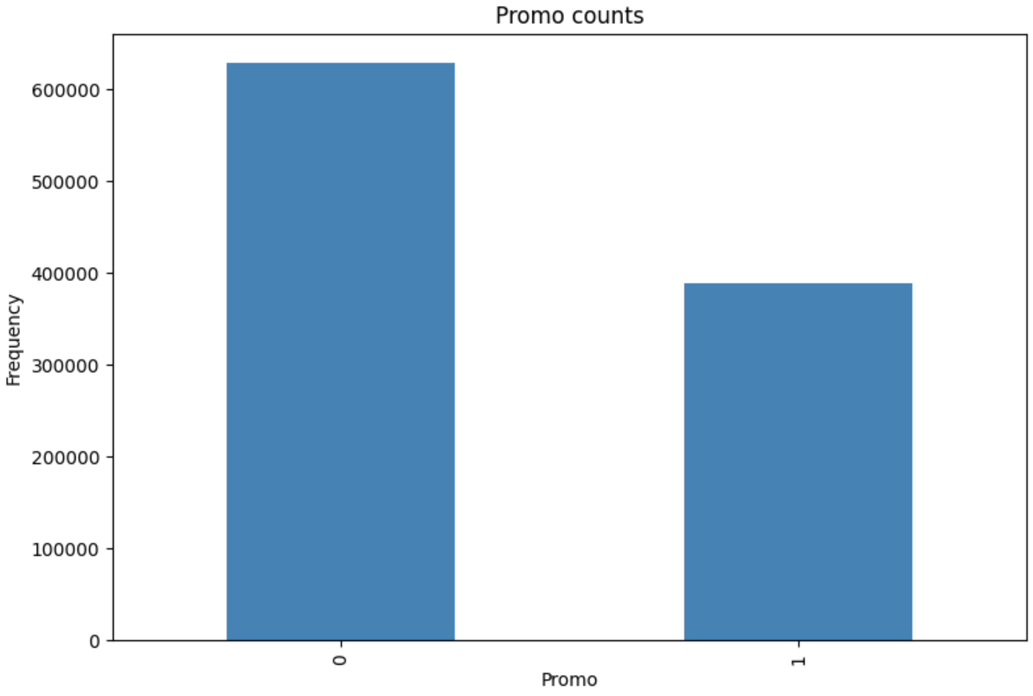
Analysis with respect to Customer

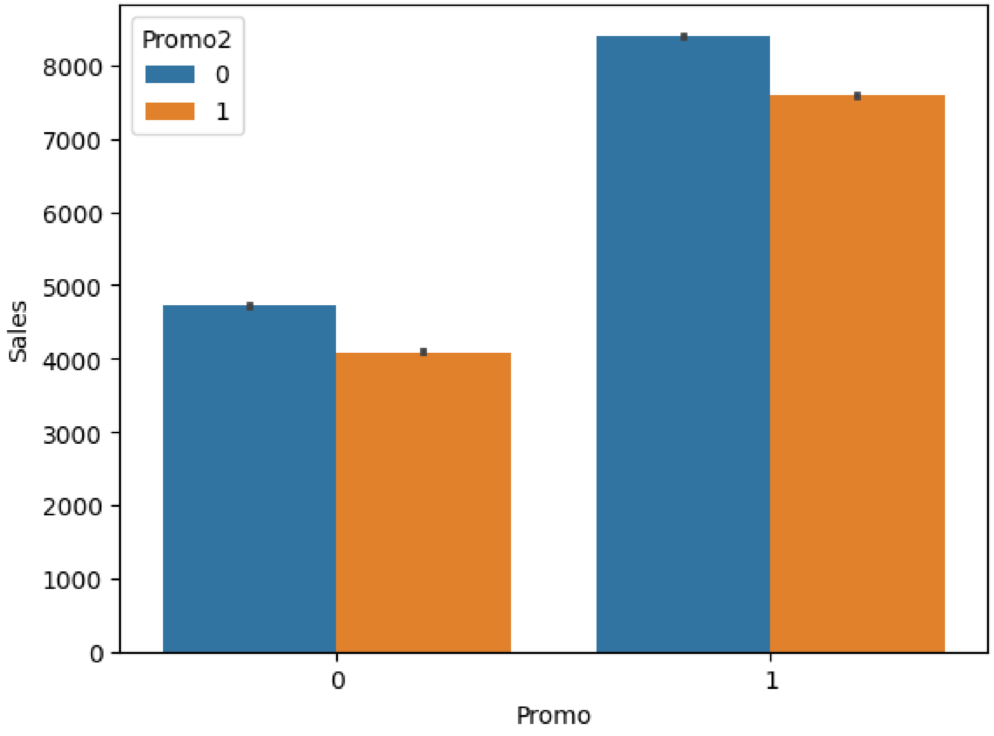




- It is evident from the scatter plot that there is a direct positive correlation between customer and sales.
- As discussed earlier Store type 'a' has more customer visits because there are more number of stores with type 'a'. However store type 'b' has better customer to sales ratio, because store type 'b' uses all three levels of assortment.

Analysis with respect to Promotion





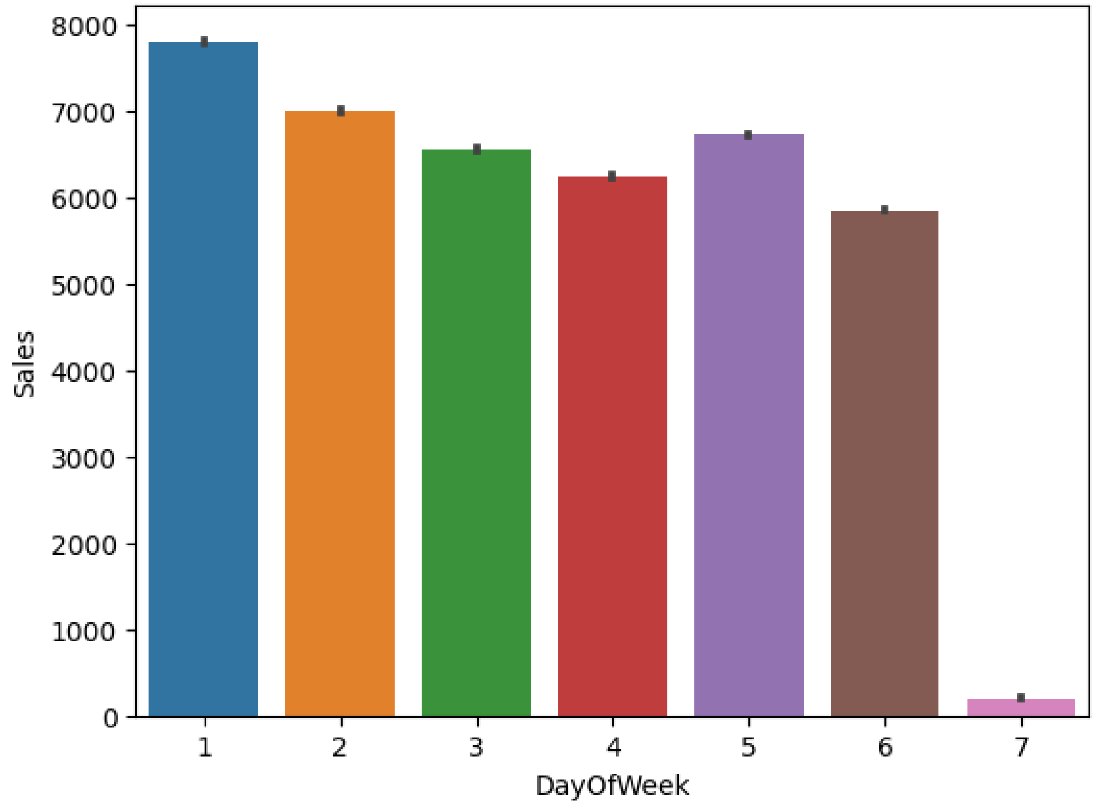
- Stores participating in Promo are comparatively less than the ones not participating.

- It is very evident from the above the graph that the sales increases by almost 44% for the stores participating in the Promo.

Hypothesis Testing

- Based on our EDA, we have defined these three hypothesis.
  - Average Sales on other days of the week is higher than 7th day of the week. (Right Tailed test)
  - Store Type and Assortment level have some impact on sales. (Two Tailed Test)
  - Average sales for stores without promotion is less than the stores with promotion. (Left Tailed Test)

Hypothesis – 1



- From EDA we found that average sales are very low on 7th day of the week which indicate that most of the stores are closed on Sunday. We have performed 2 sample t-test to find weather their is a Increase in average sales on other days of the week in compare to the 7th day of the week (Right Tailed Test).

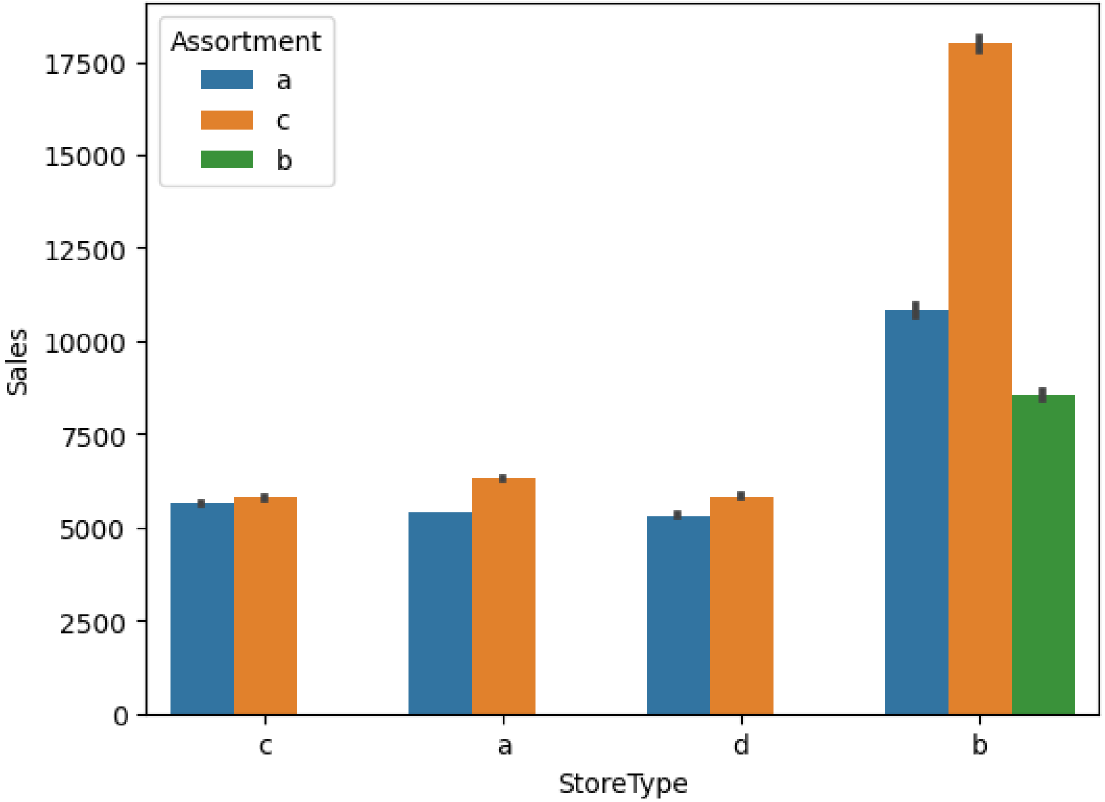
- **Null Hypothesis(H_O) :** mean of other_days - mean of seventh_day = 0

- **Alternate Hypothesis(H_A) :** mean of other_days > mean of seventh_day

- **Significance level :** alpha = 0.05

- **We got p-value = 0.00 which is less then significance level thus we can reject the null hypothesis.**

Hypothesis – 2

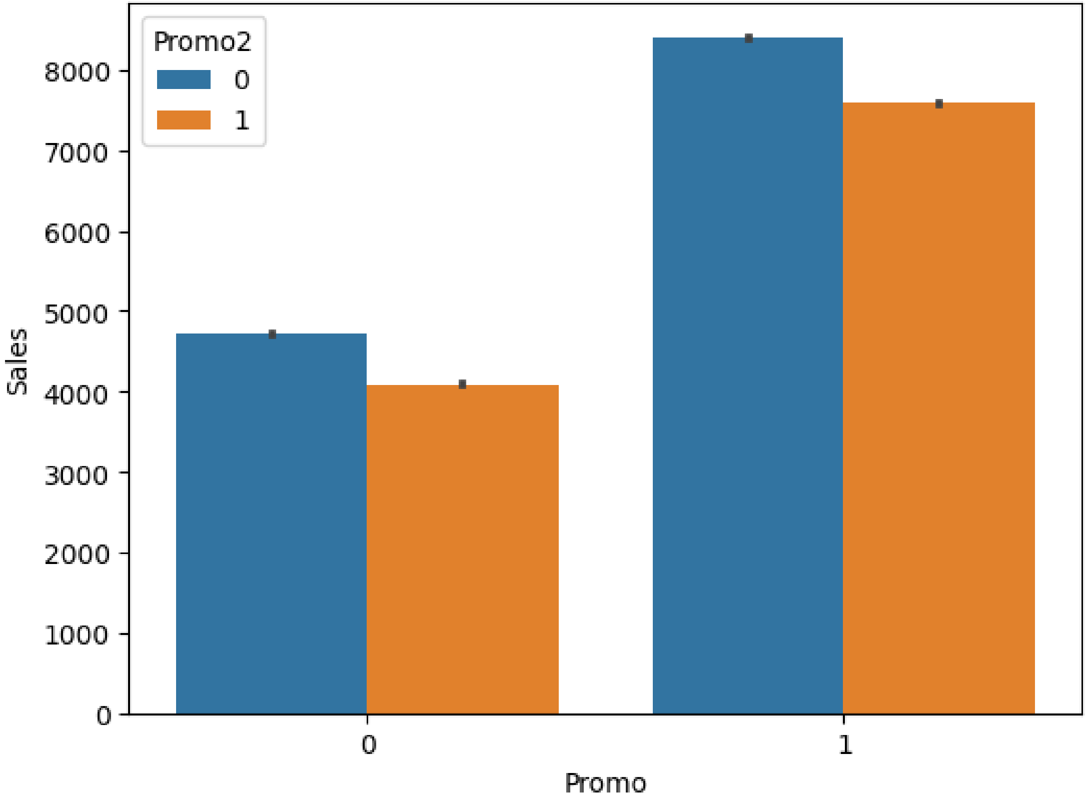


- From EDA we found that store type 'b' has highest average sales and per store revenue as it works with all three level of assortment level. We have performed a two-way ANOVA test to find out what impact store type and assortment level have on sales (Two Tailed Test).

- **Null Hypothesis(H_O) :** There is no difference in average sales for any store type.
  - **Alternate Hypothesis(H_A) :** There is a difference in average sales for any store type.
- **Null Hypothesis(H_O) :** There is no difference in average sales with any assortment level.
  - **Alternate Hypothesis(H_A) :** There is a difference in average sales with any assortment level.
- **Null Hypothesis(H_O) :** The effect of Store type on average sales does not depend on the effect of the assortment level (a.k.a. no interaction effect).
  - **Alternate Hypothesis(H_A) :** There is an interaction effect between assortment level and store type on average sales.

- **Significance level :** alpha = 0.05
- **We got p-value = 0.00 which is less then significance level thus we can reject the null hypothesis.**

Hypothesis – 3



- From EDA we found out that with promotion average sales increase by almost 44%. We'll use 2 sample t-test to check the decrease in sales for store without promotion. (Left Tailed Test).

- **Null Hypothesis(H_O) :** mean without Promo - mean with Promo = 0

- **Alternate Hypothesis(H_A) :** mean without Promo < mean with Promo

- **Significance level :** alpha = 0.05

- **We got p-value = 0.00 which is less then significance level thus we can reject the null hypothesis.**

Feature Engineering and Data Pre-Processing

- **Handling Missing Values :**
  - PromoSinceWeek, PromoSinceYear and PromoInterval had more than 50% values missing so we dropped the columns.
  - CompetitionDistance has a right skewed distribution so imputing the missing value with median.
  - Imputing the missing value of CompetitionSinceMonth and CompetitionSinceYear with mode.

- **Handling Outliers :**
  - From EDA on outliers we were able to established that the outliers are showing this behavior for the stores with promotion ‘on’ and store type 'b'. It would not be wise to treat them because the reasons behind this behavior seems fair and important from the business point of view.
  - The primary reasons for the behavior are promotion and store type 'b'.
  - If the outliers are a valid occurrence it would be wise not to treat them by deleting or manipulating them.

- **Categorical Encoding :**
  - For State Holiday data point with '0' were around 90% and other data point 'a', 'b', 'c' were around 4% each. So we converted the values, '0' to 0 and 'a','b','c' to 1.
  - We have used One Hot Encoding for StoreType and Assortment as the data is nominal and not ordinal after spliting the data for model implemntation.

- **Feature Selection :**
  - We have used VIF to check what all continuous features follow multicolinearity and select the feature accordingly.
  - We directly selected only the data points where the stores were open as closed store generate 0 sales.
  - We do not need Store and Date Column for prediction, so we dropped them directly.
  - Important Continuous features are - Customers, CompetetionDistance.
  - Important Categorical Features are - StoreType, Assortment, DayOfWeek, Promo, StateHoliday, SchoolHoliday, Month, Promo2, Competition_open_total_months.

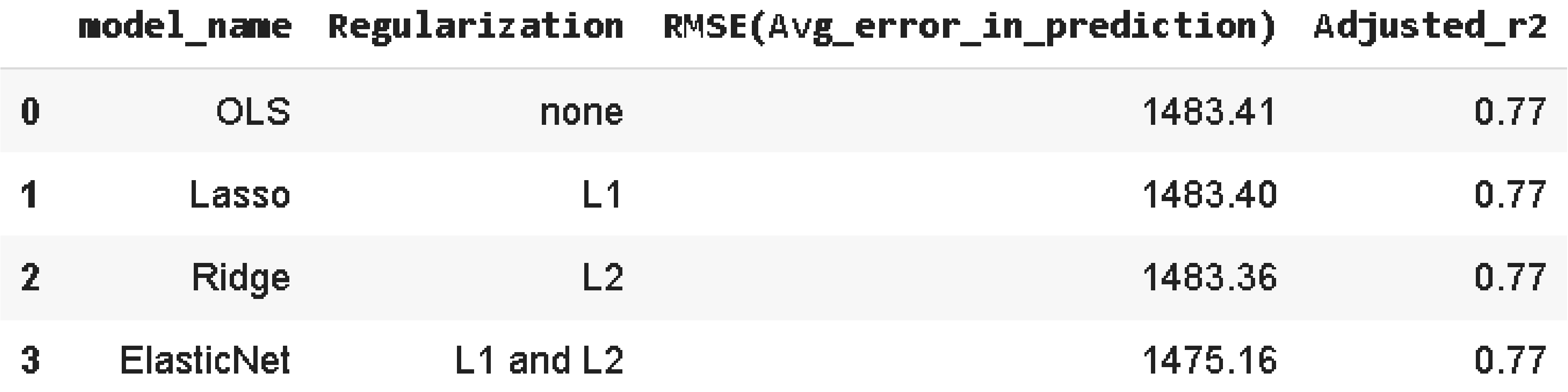
- **Data Transformation :**
  - Target Variable is right skewed so we need to transform it to normal or near normal distribution.
  - We have applied square root transformation on target variable to transform it to near normal distribution.

Model Implementation

- We have Tried implementing the following models on our dataset :
- Linear Regression
  - OLS(Ordinary Least Square)
  - Lasso Regression
  - Ridge Regression
  - Elasticnet Regression
- Polynomial Regression
  - Elasticnet Polynomial Regression
- Decision Tree
- Random Forest

Linear Regression

- We Implemented Linear Regression model and got the following results.

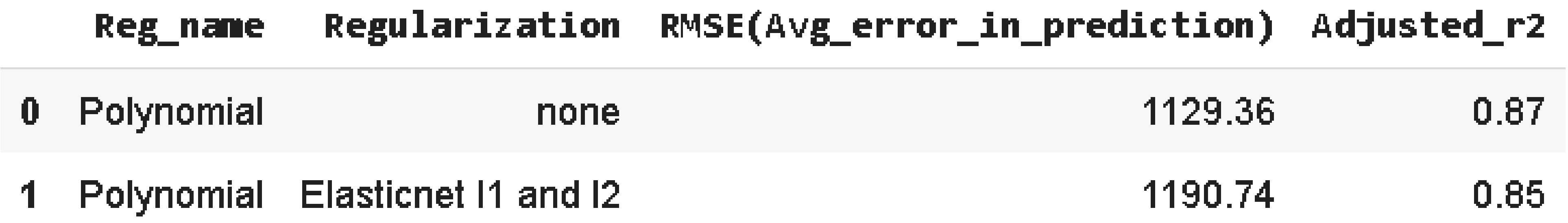


	model_name	Regularization	RMSE(Avg_error_in_prediction)	Adjusted_r2
0	OLS	none	1483.41	0.77
1	Lasso	L1	1483.40	0.77
2	Ridge	L2	1483.36	0.77
3	ElasticNet	L1 and L2	1475.16	0.77

- As this dataset has pattern such as peak days, festive seasons, etc. which are most likely considered as outliers in Linear Regression. So Linear Model may not show good results in such cases.
- So we tried to fit Polynomial Regression on the dataset in aim of getting better accuracy.

Polynomial Regression

- We Implemented Polynomial Regression model and got the following results :

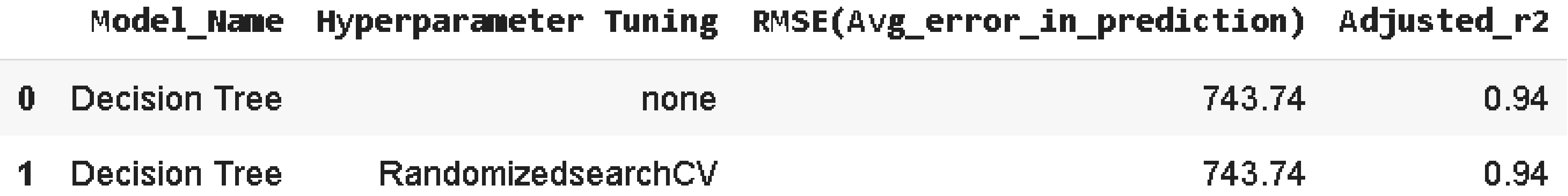


	Reg_name	Regularization	RMSE(Avg_error_in_prediction)	Adjusted_r2
0	Polynomial	none	1129.36	0.87
1	Polynomial	Elasticnet l1 and l2	1190.74	0.85

- Polynomial Regression is has better accuracy in sales prediction as compare to linear regression model on this dataset.
- Average error has decreased from 1475.16 to 1129.36.
- And variance explained by Polynomial Regression is high i.e. 87 % which is 10 % higher than Linear Regression.
- However Tree based models are robust with outliers, so we have implemented tree based model to this dataset and compared the results.

Decision Tree

- We Implemented Decision Tree Model and got following results :

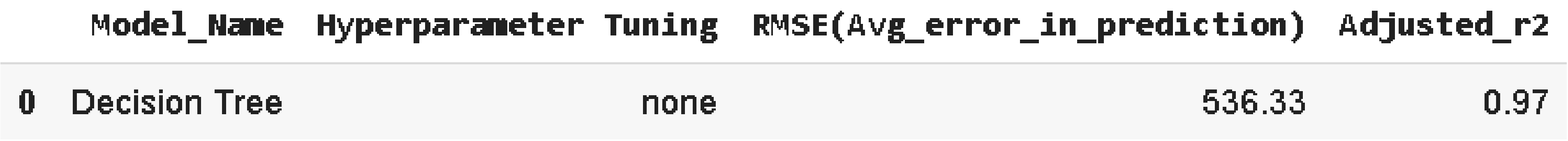


	Model_Name	Hyperparameter Tuning	RMSE(Avg_error_in_prediction)	Adjusted_r2
0	Decision Tree	none	743.74	0.94
1	Decision Tree	RandomizedsearchCV	743.74	0.94

- Average error decreased from 1129.36 to 743.74 and decision tree model is able to explain 94 % of the variance.

Random Forest

- We Implemented Random Forest model and got the following results :



	Model_Name	Hyperparameter Tuning	RMSE(Avg_error_in_prediction)	Adjusted_r2
0	Decision Tree	none	536.33	0.97

- In Compare to every other model Random forest fits the best.
- Average Error is decreased from 743.74 to 536.33 and Random Forest is able to explain 97 %   of the variance.

Evaluation Metrics

- Different metrics may be more suitable for different purposes and audiences. For example, some metrics may focus on the accuracy of predictions, while others may focus on the variability explained by the model. The Metrics used in the Project are.

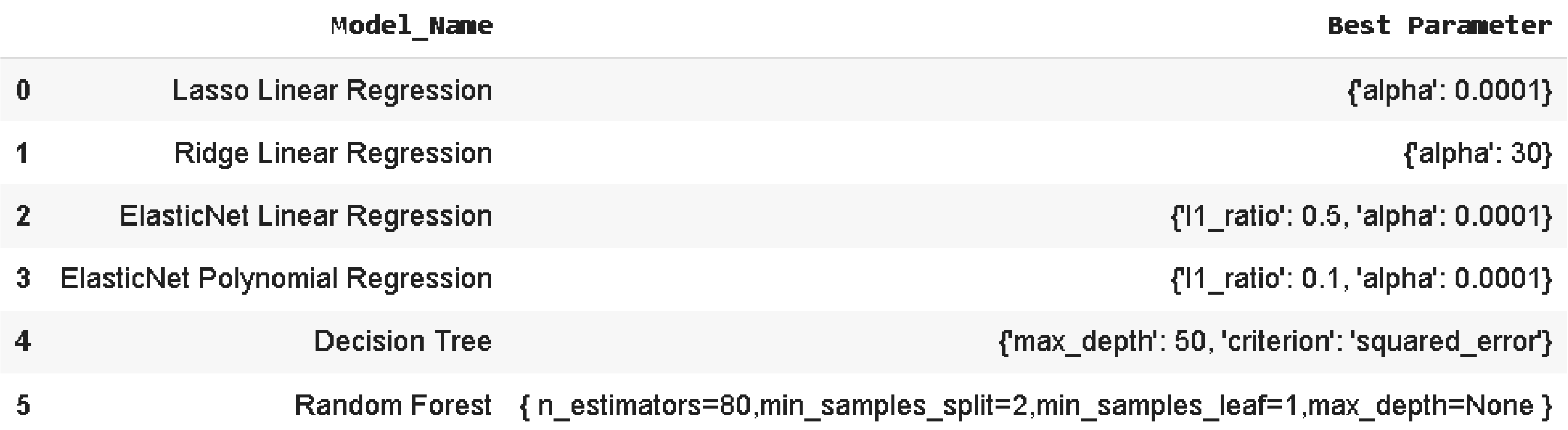
  - **RMSE Root Mean Squared Error :** It measures average error there is in the predictions. 536 is the average prediction error for a sale

  - **Adjusted r squared :** it measures the proportion of the variance in the response variable that is explained by the predictor variables, adjusted for the number of predictors. It is a modified version of r2 that takes into account the number of predictors and the sample size. It penalizes the model for adding irrelevant predictors

- **Best Results :** Random Forest model has the Lowest values for RMSE i.e. 536 and Highest value for Adjusted r squared, i.e. 0.97 which means 97 %   variability can be explained using the model.

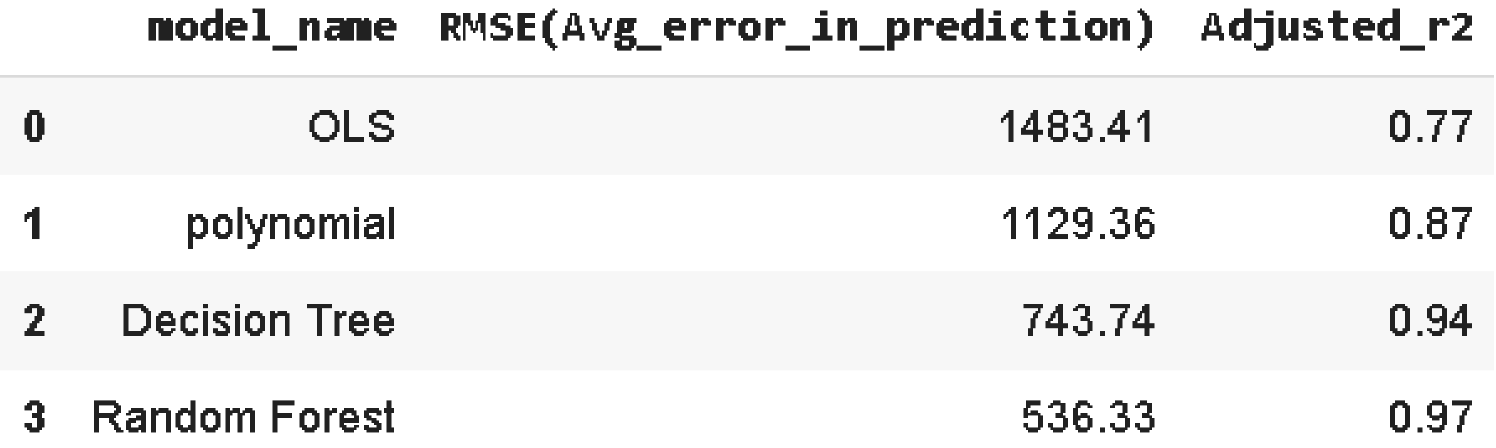
Hyperparameter Tuning

- We have used RandomizedSearchCV for hyperparameter tuning.
- RandomizedSearchCV solves the drawbacks of GridSearchCV, as it goes through only a fixed number of hyperparameter settings. It moves within the grid in a random fashion to find the best set hyperparameters.
- This approach reduces unnecessary computation.
- Here are is the chart for best parameters obtained by RandomizedSearchCV for different models.



	Model_Name	Best Parameter
0	Lasso Linear Regression	{'alpha': 0.0001}
1	Ridge Linear Regression	{'alpha': 30}
2	ElasticNet Linear Regression	{'l1_ratio': 0.5, 'alpha': 0.0001}
3	ElasticNet Polynomial Regression	{'l1_ratio': 0.1, 'alpha': 0.0001}
4	Decision Tree	{'max_depth': 50, 'criterion': 'squared_error'}
5	Random Forest	{ n_estimators=80,min_samples_split=2,min_samples_leaf=1,max_depth=None }

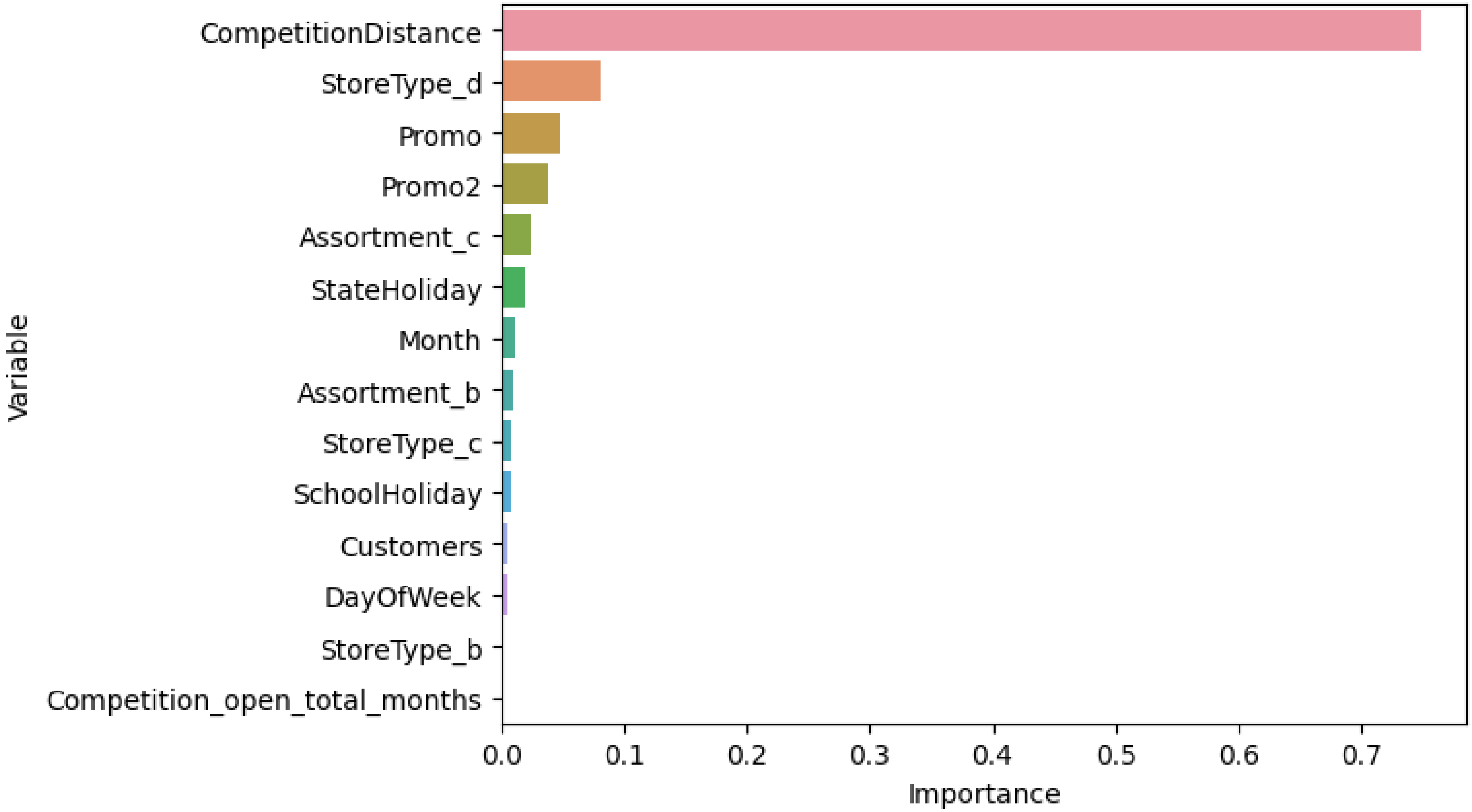
Model Evaluation



	model_name	RMSE(Avg_error_in_prediction)	Adjusted_r2
0	OLS	1483.41	0.77
1	polynomial	1129.36	0.87
2	Decision Tree	743.74	0.94
3	Random Forest	536.33	0.97

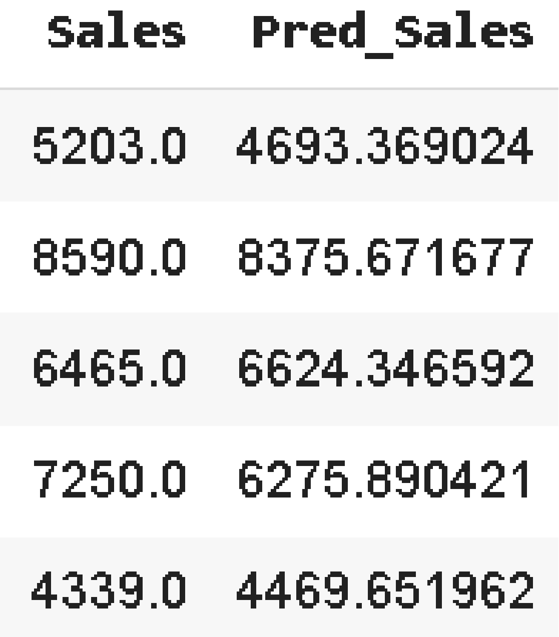
- Out of all models Random Forest best fit the data.
- Average error is least i.e. is 536.33 and variance explained by Random Forest model is the highest i.e. 97 % .
- So using Random Forest for sales prediction can predict the sales more accurately as compare to other models.

Feature Importance



Sales Prediction

- Here is the comparison between actual sales and predicted sales after implementing Random Forest model.



Sales	Pred_Sales
5203.0	4693.369024
8590.0	8375.671677
6465.0	6624.346592
7250.0	6275.890421
4339.0	4469.651962

Conclusion

- With Promotion there is an increase in sales by almost 44 % . Promotions also have a positive impact on customer. However there are comparatively less number of stores participating in Promotion. So more Stores should be encouraged to participate in Promotion.
- Store type ‘b’, in spite of being less in number has highest average sales and per store revenue, as it works with all three assortment levels and is also open on Sunday(7th day of the week). There should be more stores with type ‘b’ or the same strategies should be used for other store type as store type ‘b’.
- Talking about competition distance store densely located near each other experience more sales than the ones that are located further.
- Random Forest model fit best on the data, with least average error i.e. 536 and variance explained by Random forest is highest i.e. 97 %   as compare to other models. Deploying Random Forest can give more accurate sales prediction.