Rossman Retail Sales Prediction

Prepared by:
Saumya Dash & Kumar Abhinav
Data science trainees,
AlmaBetter, Bangalore

Abstract:

Sales forecasting is the process of estimating demand for or sales of a specific product over a specific time period. Businesses use sales forecasts to determine how much revenue they will generate in a given time period, allowing them to create powerful and strategic business plans. Budgets, hiring, incentives, goals, acquisitions, and other growth plans are all influenced by the revenue the company expects to generate in the coming months, and for these plans to be as effective as they are expected to be, these forecasts must also be as good.

Keywords: EDA, Correlation, Decision Tree, Random Forest, Regression, Forecasting

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.

we 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.

Introduction:

The popularity of a product fluctuates from time to time. No company can plan for financial growth unless it accurately assesses customer interest and future demand for products. Sales forecasting is the process of estimating demand for or sales of a specific product over a specific time period. It is critical to obtain a good dataset in order to make an accurate sales forecast. Forecasts are heavily reliant on previous sales records, trends, and patterns observed in a specific store. The variations could be due to a variety of factors.

From a business standpoint, these sales forecasts are done on a regular basis to improve their sales forecasting models because they have a direct impact on their decision making process, goals, plans, and growth strategies.

Machine learning models are created and compared in this Retail Sales Prediction to predict sales of these 1115 drug stores across the European market. In addition, an effort has been made to analyse and identify all of the features that contribute to higher sales and those that contribute to lower sales, so that improvement plans can be developed.

Approach:

The approach followed here is to first check thesanctity of the data and then understand the features involved. The events followed were in our approach:

- Understanding the business problemand the datasets
- **Data cleaning and preprocessing-** finding null values and imputing them with appropriate values.
  Converting categorical values into appropriate data types and merging the datasets provided to get a final dataset towork upon.
- **Exploratory data analysis-** of categorical and continuous variables against our target variable.
- **Data manipulation-** feature selection and engineering, feature scaling, outlier detection and treatment and encoding categorical features.
- **Modeling-** The baseline model- Decision tree was chosen considering our features were mostly categorical with few having continuous importance.
- **Model Performance and Evaluation**
- **Store wise Sales Predictions**
- **Conclusion and Recommendations**

Data Exploration

First step involved is understanding the data and getting answers to some basic questions like, "What is the data about?" What is the number of rows or observations in it? How many features does it have? What are the various data types? Are there any values missing? And anything else that might be relevant or helpful to our investigation. Before proceeding, let us first understand the dataset and the terms involved.

The first csv file in our dataset i.e Rossman sales dataset contains historical data with 1017209 rows or observations and 9 columns with no null values. The second dataset contained additional information about the stores, with 1115 rows and 10 columns, as well as many missing values in a few columns. The data types were integers , float and object data types

The details of attributes of Rossman sales data set is given below:

Columns	Description
ID	an Id that represents a (Store, Date) duple within the set
DayOfWeek	Day of the week
Date	Date of the Sales
Sales	The turnover for any given day (this is what you are predicting)
Customers	The number of customers on a given day
Open	An indicator for whether the store was open: 0 = closed, 1 = open
Promo	indicates whether a store is running a promo on that day
StateHoliday	indicates a state holiday. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
SchoolHoliday	indicates if the (Store,Date) was affected by the closure of public schools

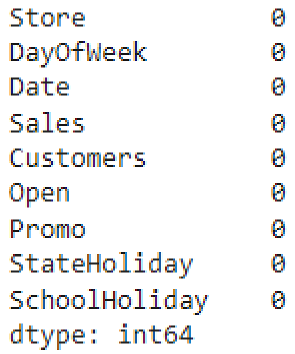
Similarly for the user review dataset details of its column are given below

Columns	Description
App	The app name.
Translated_Review	Review text in English.
Sentiment	Sentiment of the review, which can be positive, neutral, or negative.
Sentiment_Polarity	Sentiment in numerical form, ranging from -1.00 to 1.00.

Sentiment_Subjectivity	Measure of the expression of opinions, evaluations, feelings, and speculations

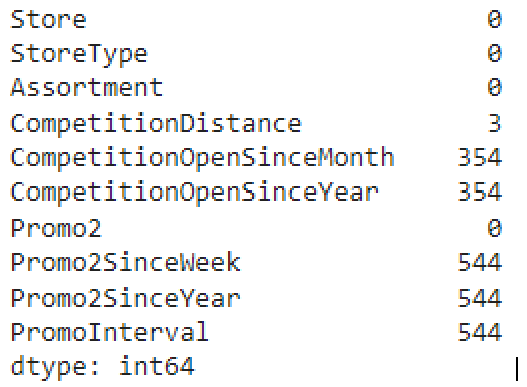
Handling missing values is an important skill in the data analysis process. If there are very few missing values compared to the size of the dataset, we may choose to drop rows that have missing values. Otherwise, it is better to replace them with appropriate values. It is necessary to check and handle these values before feeding it to the models, so as to obtain good insights on what the data is trying to say and make great characterization and predictions which will in turn help improve the business's growth.

The historical record dataset has no null values(sales_df) :



```
Store            0
DayOfWeek        0
Date             0
Sales            0
Customers        0
Open             0
Promo            0
StateHoliday     0
SchoolHoliday    0
dtype: int64
```

While Store dataset have some null values which we can see from the table illustrated below:



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

The dataset had a lot of nulls in the following columns:

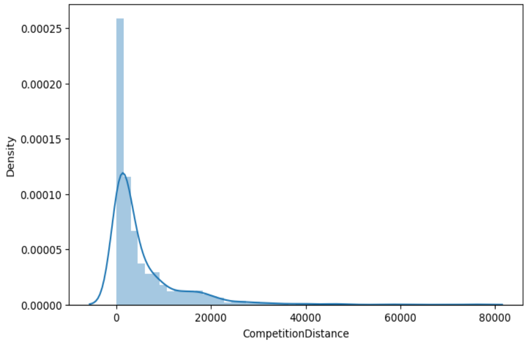
- CompetitionDistance
- CompetitionOpenSinceMonth
- CompetitionOpenSinceYear
- Promo2SinceWeek
- Promo2SinceYear
- PromoInterval

We treated each of the columns one by one ;

‘CompetitionDistance’ –

Competition Distance is the distance in meters to the nearest competitor store.

The Competition Distance distribution plot shows the distances at which generally the stores are opened



Majority of the CompetitionDistance values appear to be on the left, with the distribution skewed to the right. Because median is more resistant to outlier effects, it was used to impute null values.

Right skewed distributions occur when the long tail is on the right side of the distribution, also known as a positive skewed distribution, implying that there are positive outliers far along in the distribution that influence the mean.

The majority of the CompetitionDistance values in the column appear to be between 0 and 10 kilometres. As a result, in an asymmetrical distribution, the longer tail pulls the mean away from the most common values. The mean outnumbers the median. The mean overestimates the most common values in the distribution, so the median is used in this case. The median is also more resistant to outlier effects, so it is used to impute missing values in this feature.

- **CompetitionOpenSinceMonth-** gives theapproximate month of the time the nearest competitor was opened. The mode of the column is used to impute the missing values in the column as it gives the most occurring month.
- **CompetitionOpenSinceYear**-gives the approximate year of the time the nearest competitor was opened. The mode of thecolumn is used to impute the missing values in the column as it gives the mostoccurring month.
- **Promo2SinceWeek**, **Promo2SinceYear** and **PromoInterval** are NaN wherever Promo2 is 0 or False as can be seen in the first look of the dataset. They are replaced with0.

Lastly before proceeding further, the two datasetswere merged on the common column of ‘Store’ to get everything together for the analysis.

Exploratory Data Analysis:

Exploratory data analysis is a crucial part of data analysis. It involves exploring and analyzing the dataset given to find out patterns, trends and conclusions to make better decisions related to the data, often using statistical graphics and other data visualization tools to summarize the results. The visualization tools involved in the investigation are python libraries- matplotlib and seaborn.

The goal here is to explore the relationships of different variables with ‘Sales’ to see what factorsmight be contributing to the high and low sales numbers.

Approach:

The dataset contains two types of features: categorical variables and noncategorical variables.

Categorical - A categorical variable is one that can take on one of a limited, and usually fixed, number of possible values, assigning the observation to a specific category.

Non Categorical - A non categorical or continuous variable is one whose value is obtained through measurement, i.e., one with an uncountable set of values.

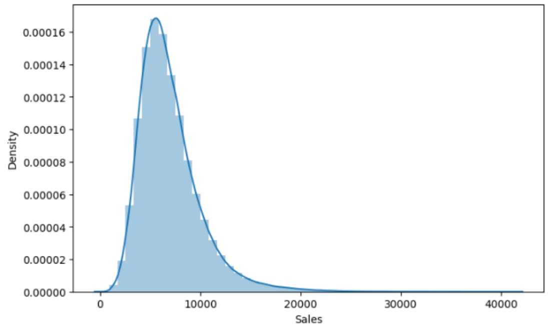
They are both examined separately. Categorical data is typically analyzed using count plots and barplots in relation to the target variable, and this is done here as well. Numeric or continuous variables, on the other hand, were analyzed using distribution plots, box plots, and scatterplots to gain useful insights.

Univariate Analysis:

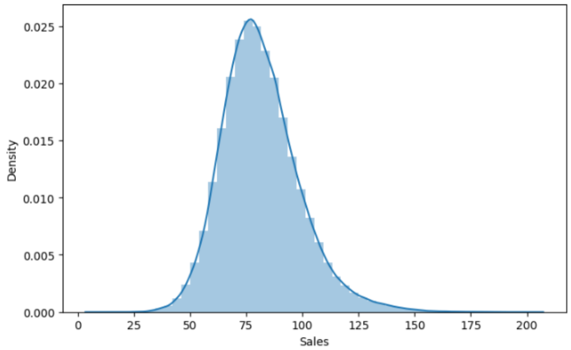
Continuous Features:

We have continuous features as Sales, Customers and CompetitionDistance. Lets check the distribution of each of them one by one.

Sales:

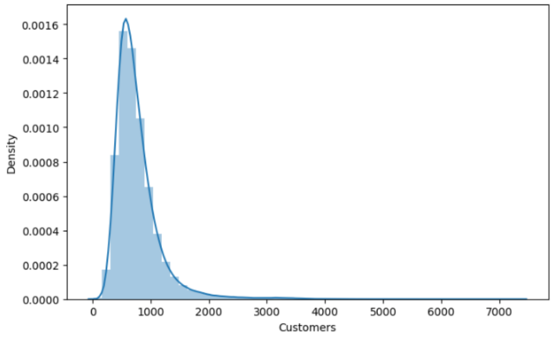


We can observe from the above plot that Sales distribution is right skewed we can apply square root transformation to treat it and make it best fit for modelling.

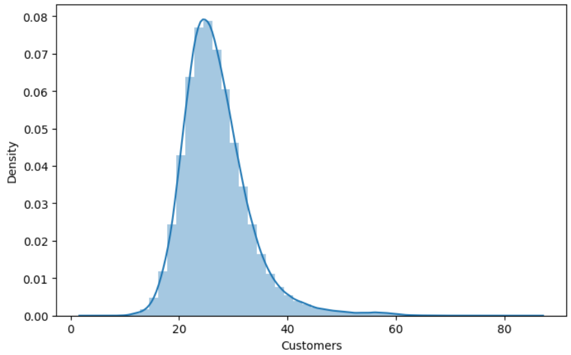


After applying the transformation we can observe that it is normally distributed

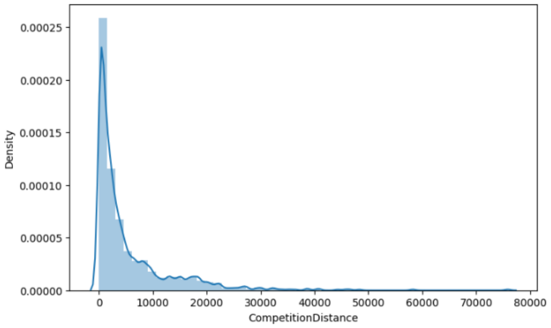
Customers:



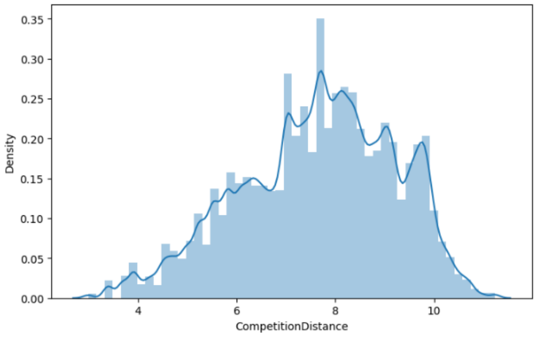
Similarly we can apply square root transformation in customer column to make it normally distributed.



CompetitionDistance:



Spread of CompetitionDistance is also right skewed as well as there is a huge bump at 0. We applied log transformation for this case

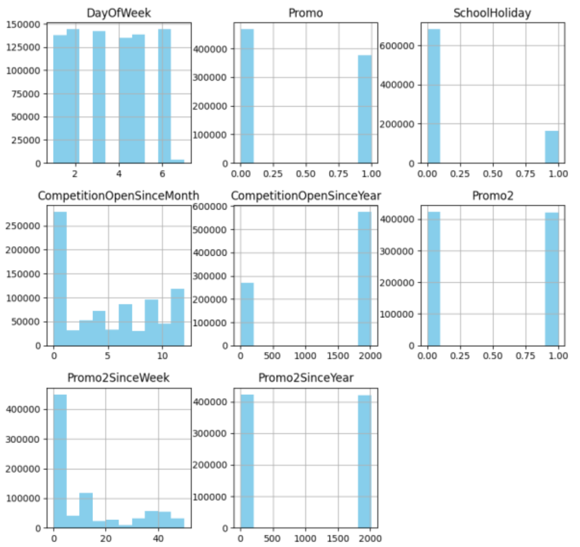


Categorical Features:

For this we have plotted histograms to check the frequency :
List of categorical features for which we have plotted histograms:

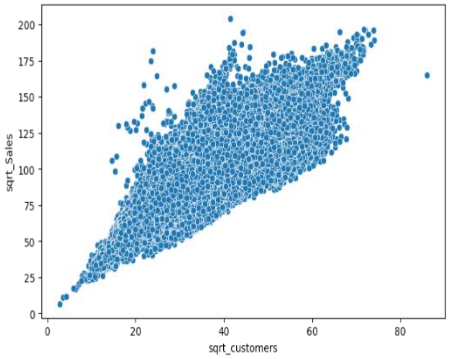
- DayOfWeek
- Promo
- SchoolHoliday
- CompetitionOpenSinceMonth
- CompetitionOpenSinceYear

- Promo2
- Promo2SinceWeek
- Promo2SinceYear



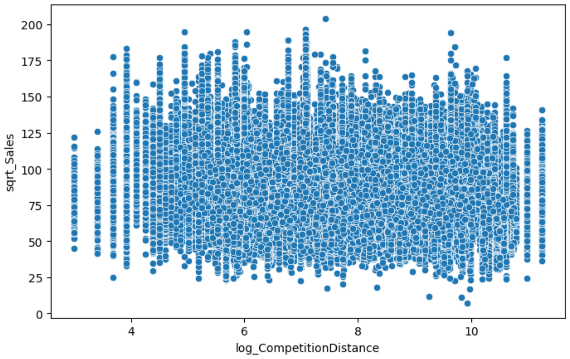
Bivariate Analysis:

Sales vs Customer



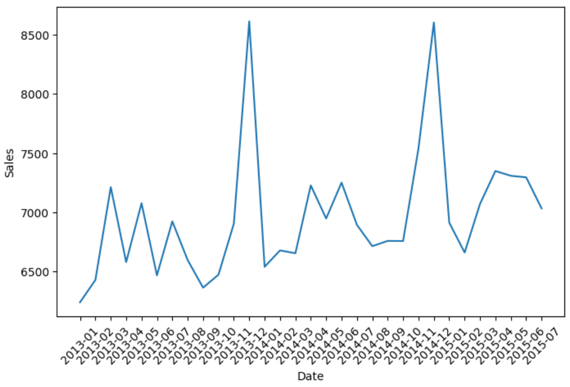
We have applied square root transformation to both Sales and Customer columns and then we plotted scatter plot between them we can observe that spread is showing the linear relation between the two variables , at the same time spread is fairly homoscedastic **.**

Sales vs CustomerDistance



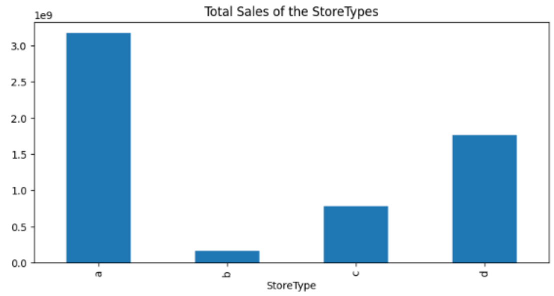
We have applied log transformation to CompetitionDistance and then we plotted scatter plot between Sales and CompetitionDistance. We can observe that spread is uniform and linear.

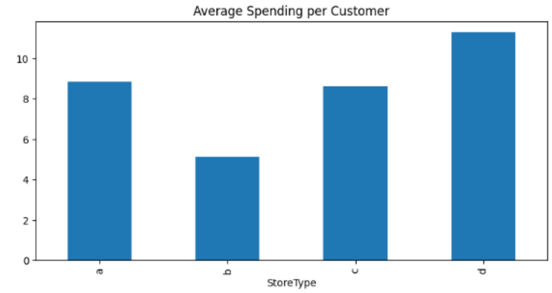
Monthly Sales



We have plotted line plot between months and Sales , we can observe that October month of both 2013 and 2013 has maximum sales. Hence we can say that October month is the best time in terms of sales.

Sales vs StoreType

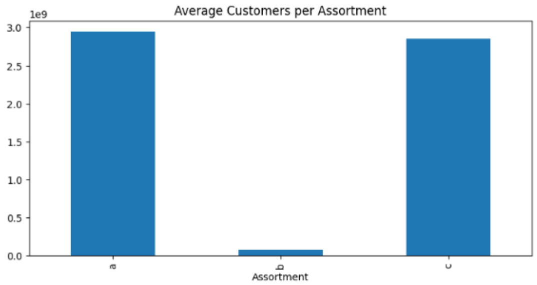


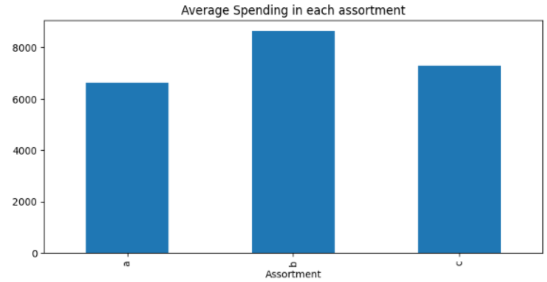


We have plotted bar plot for Total Sales vs StoreType and Average spending per customer vs StoreType to check the variation.

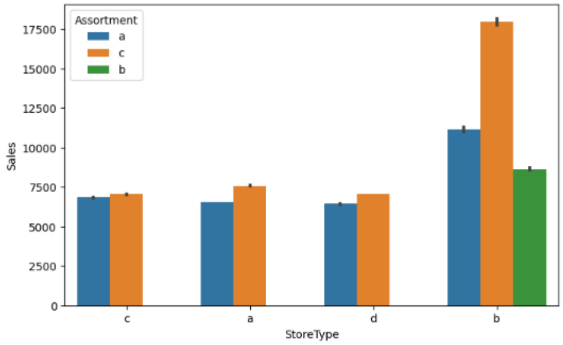
We can observe that StoreType ‘a’ has maximum total sales but StoreType ‘d’ has maximum average spending per customer.

Sales vs Assortment



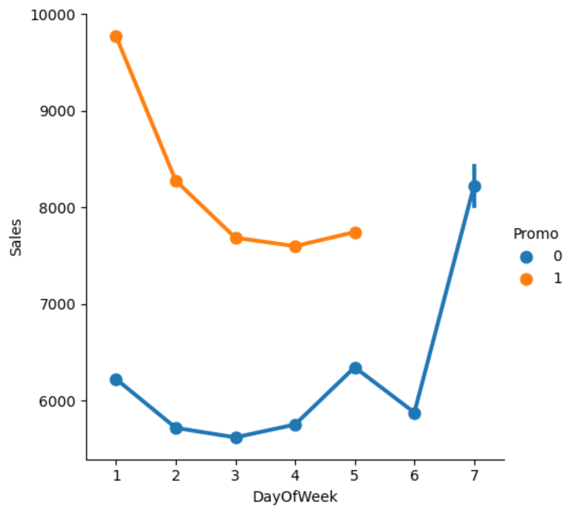


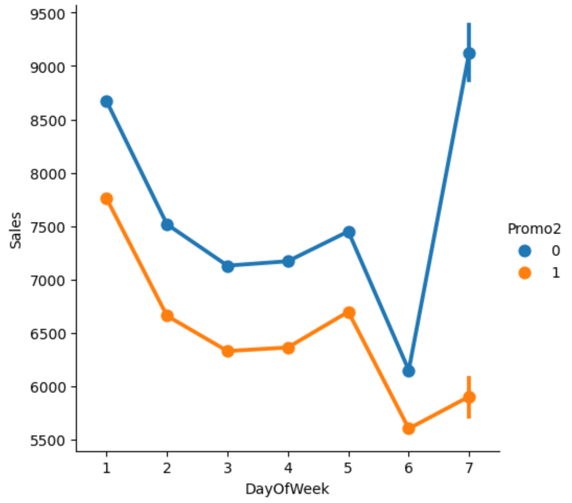
We have plotted bar plot for Total Sales vs Assortment and Average Spending in each assortment. We can observe from the that assortment a and c has maximum sales but in terms of maximum average spending per customer assortment b is performing best.



The above 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 were high for store type b stores

Sales vs DayofWeek (with or without promo)

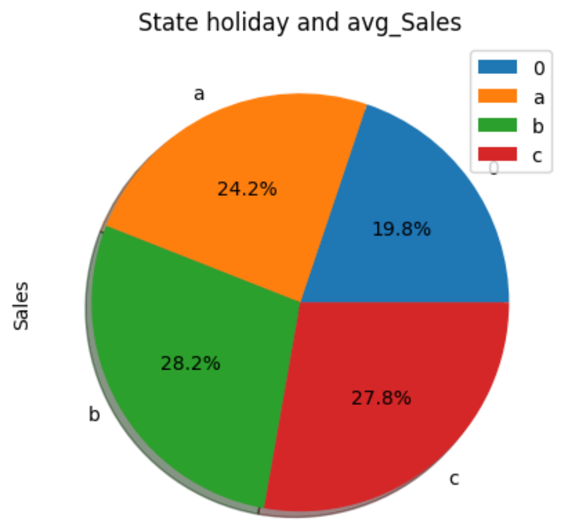


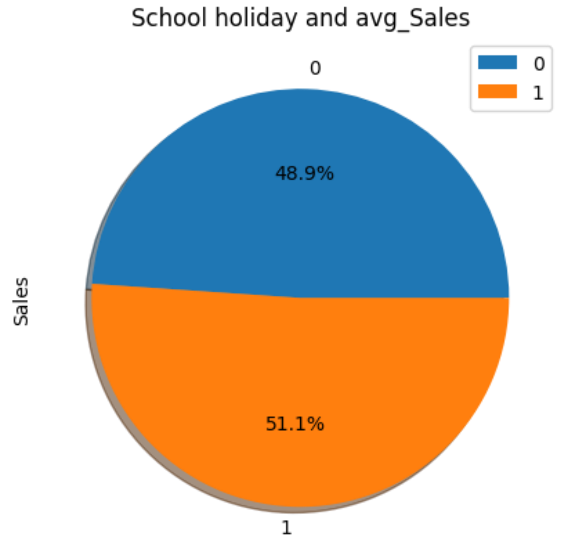


We have plotted factor plot for Sales vs dayofweek with Promo in effect and not in effect.

We can observe from the above plot that, there is no promotion in the weekend. However, the sales are very high. If promo offers are to be given on weekends definately the sales are going to skyrocket.

Sales vs Holidays



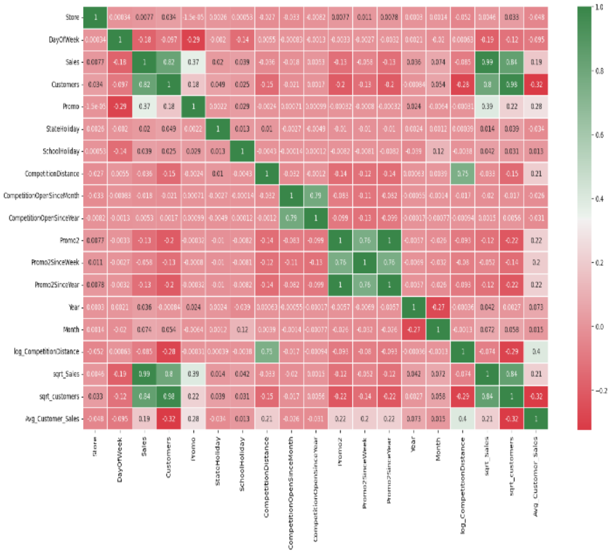


From above two pie charts it is Clearly evident that on holiday avg sales are better on holidays as compared to non holiday.

Correlation:

Correlation is a statistical term that describes how two variables move in relation to one another. A perfect positive correlation is one with a correlation coefficient of exactly one. This means that when one variable moves, either up or down, the other follows suit. A perfect negative correlation indicates that two variables move in opposite directions, whereas a zero correlation indicates that there is no linear relationship at all.

By checking the correlation the factors affecting sales can be figured out.



- Day of the week has a negative correlation indicating low sales as the weekends, and promo, customers and open has positive correlation.
- CompetitionDistance showing negative correlation suggests that as the distance increases sales reduce, which was also observed through the scatterplot earlier.
- There's multicollinearity involved in the dataset as well. The features telling the same story like Promo2, Promo2 since week and year are showing multicollinearity.

Data Manipulation:

Data manipulation entails changing and manipulating our dataset before feeding it to various regression machine learning models. This includes retaining important features, treating outliers, scaling features, and, if necessary, creating dummy variables.

Feature Engineering:

- Some stores were closed due to refurbishment and some on account of week off or holidays. Those stores on those dates generated zero sales and hence removing the rows was important to avoid confusion by the algorithms and then removing the feature altogether because it wasn’t providing any value in prediction of the sales.

- StateHoliday which were categorized into three types of different holidays were binary encoded with 1 and 0 which meant that either holiday or not.
- One hot encoding- The integer encoding is insufficient for categorical variables where no such ordinal relationship exists. Although we have categorical data integers encoded, assuming a natural order and allowing this data to the model may result in poor performance.

  Many of the features, such as DayofWeek, StoreType, and Assortments, were categorical in nature and needed to be one hot encoded in order to function.

Feature Scaling:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is done to prevent biased nature of machine learning algorithms towards features with greater values and scale. The two techniquesare:

Normalization: is a scaling technique in which values are shifted and rescaled so that they end upranging between 0 and 1. It is also known as
Min-Max scaling. [0,1]

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization: is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and theresultant distribution has a unit standard deviation. [-1,1]

$$X' = \frac{X - \mu}{\sigma}$$

Normalization of the continuous variables wasdone further.

Modeling:

Factors affecting in choosing the model:

Determining which algorithm to use depends on many factors like the problem statement and the kind of output you want, type and size of the data, the available computational time, number of features, and observations in the data, to name a few.

The dataset used in this analysis has:

- A multivariate time series relation with sales and hence a linear relationship cannot be assumed in this analysis. This kind of dataset has patterns such as peak days, festive seasons etc which would most likely be considered as outliers in simple linear regression.
- Having X columns with 30% continuous and 70% categorical features. Businessesprefer the model to be interpretable in nature and decision based algorithms work better with categorical data.

Train-Test Split:

In machine learning, train/test split splits the data randomly, as there's no dependence from one observation to the other. That's not the casewith time series data. Here, it's important to use values at the rear of the dataset for testing and everything else for training.
We kept 30% of the data for testing rest of the data we utilized for training different models.

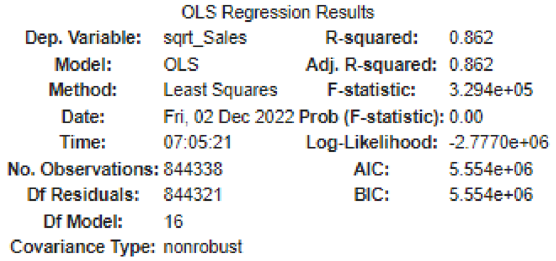
Linear Regression(OLS)

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Assumptions:

- Linear relationship between independent and dependent variables.
- Normality of the residuals.
- Homoscedasticity
- No or little multicollinearity.

OLS summary using statsmodel:



OLS Regression Results

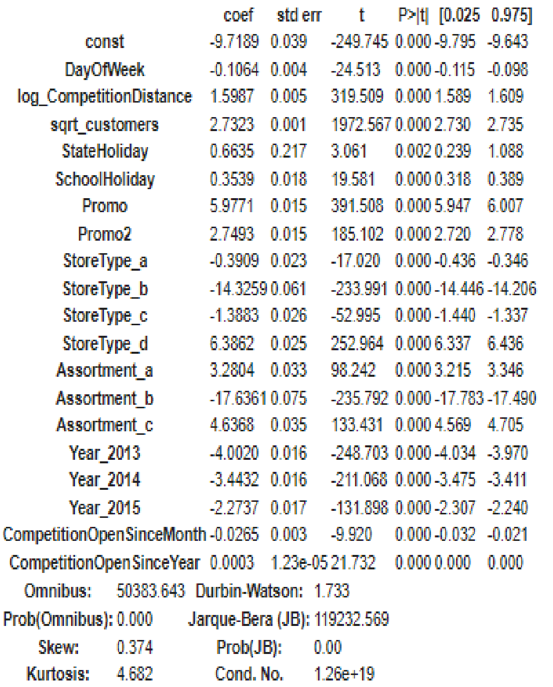
Dep. Variable:	sqrt_Sales	R-squared:	0.862
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	3.294e+05
Date:	Fri, 02 Dec 2022	Prob (F-statistic):	0.00
Time:	07:05:21	Log-Likelihood:	-2.7770e+06
No. Observations:	844338	AIC:	5.554e+06
Df Residuals:	844321	BIC:	5.554e+06
Df Model:	16		
Covariance Type:	nonrobust		

We can observe from the model summary that R squared and adjusted R squared are very close to means that means our model is performing well.

R Squared (R^2)- R2 score is a metric that tells the performance of your model,not the loss in an absolute sense that howwell did your model perform. Hence, R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit. It's value ranges from0 to 1. It can be negative if the model is performing worse than the base.

Adjusted R Squared- The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes thatwhile adding more data variance of dataincreases. Adjusted R^2 is adjusted for this disadvantage and shows the real value.

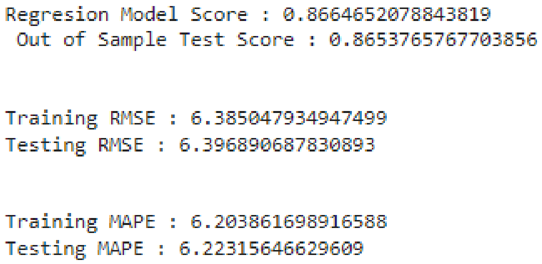
Coefficients and their p-value:



	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.7189	0.039	-249.745	0.000	-9.795	-9.643
DayOfWeek	-0.1064	0.004	-24.513	0.000	-0.115	-0.098
log_CompetitionDistance	1.5987	0.005	319.509	0.000	1.589	1.609
sqrt_customers	2.7323	0.001	1972.567	0.000	2.730	2.735
StateHoliday	0.6635	0.217	3.061	0.002	0.239	1.088
SchoolHoliday	0.3539	0.018	19.581	0.000	0.318	0.389
Promo	5.9771	0.015	391.508	0.000	5.947	6.007
Promo2	2.7493	0.015	185.102	0.000	2.720	2.778
StoreType_a	-0.3909	0.023	-17.020	0.000	-0.436	-0.346
StoreType_b	-14.3259	0.061	-233.991	0.000	-14.446	-14.206
StoreType_c	-1.3883	0.026	-52.995	0.000	-1.440	-1.337
StoreType_d	6.3862	0.025	252.964	0.000	6.337	6.436
Assortment_a	3.2804	0.033	98.242	0.000	3.215	3.346
Assortment_b	-17.6361	0.075	-235.792	0.000	-17.783	-17.490
Assortment_c	4.6368	0.035	133.431	0.000	4.569	4.705
Year_2013	-4.0020	0.016	-248.703	0.000	-4.034	-3.970
Year_2014	-3.4432	0.016	-211.068	0.000	-3.475	-3.411
Year_2015	-2.2737	0.017	-131.898	0.000	-2.307	-2.240
CompetitionOpenSinceMonth	-0.0265	0.003	-9.920	0.000	-0.032	-0.021
CompetitionOpenSinceYear	0.0003	1.23e-05	21.732	0.000	0.000	0.000

Omnibus:	50383.643	Durbin-Watson:	1.733
Prob(Omnibus):	0.000	Jarque-Bera (JB):	119232.569
Skew:	0.374	Prob(JB):	0.00
Kurtosis:	4.682	Cond. No.	1.26e+19

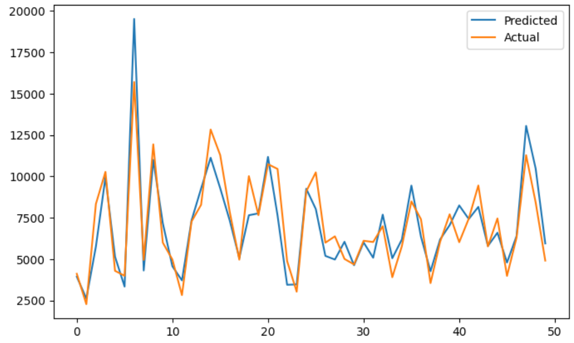
Model Performance:



```
Regresion Model Score : 0.8664652078843819
 Out of Sample Test Score : 0.8653765767703856


Training RMSE : 6.385047934947499
Testing RMSE : 6.396890687830893


Training MAPE : 6.203861698916588
Testing MAPE : 6.22315646629609
```

We can see that when we fitted the model in linear regression we came up 86% of both training and testing accuracy. RMSE and MAPE also almost equal for both training and testing dataset .But at the same time accuracy of prediction can be improved using other models . Lets check the fit in other models one by one.

Here RMSE is Root mean squared error and MAPE is mean absolute percentage error.



LASSO – LARS Regression:

Lasso regression is a type of **linear regression** that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

The acronym “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.

Lasso solutions are quadratic programming problems, which are best solved with software (like Matlab). The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

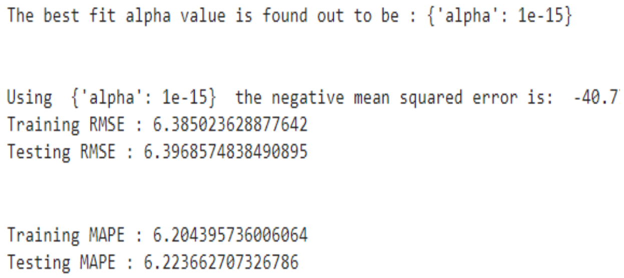
Which is the same as minimizing the sum of squares with constraint Σ $|B_j \leq s$ (Σ = summation notation). Some of the βs are shrunk to exactly zero, resulting in a regression model that’s easier to interpret.

Hyperparameter tuning:

A **tuning parameter**, α controls the strength of the L1 penalty. α is basically the amount of shrinkage:

- When α = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.
- As α increases, more and more coefficients are set to zero and eliminated (theoretically, when α = ∞, *all* coefficients are eliminated).
- As α increases, bias increases.
- As α decreases , variacnce increases
- If an intercept is included in the model, it is usually left unchanged.

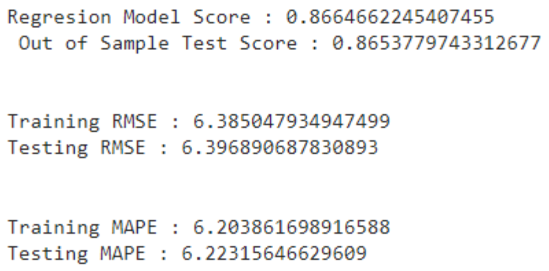
When we implemented gridserchcv algorithm to our model we came up with following outcomes:



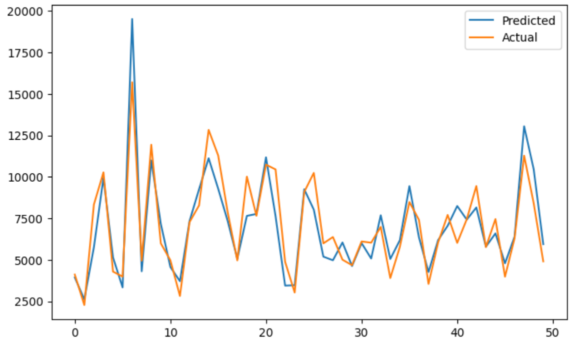
```
The best fit alpha value is found out to be : {'alpha': 1e-15}


Using  {'alpha': 1e-15}  the negative mean squared error is:  -40.7
Training RMSE : 6.385023628877642
Testing RMSE : 6.3968574838490895


Training MAPE : 6.204395736006064
Testing MAPE : 6.223662707326786
```

After using appropriate value of hyperparameter we got improved model from earlier linear regression model.

Model performance:



```
Regresion Model Score : 0.8664662245407455
 Out of Sample Test Score : 0.8653779743312677


Training RMSE : 6.385047934947499
Testing RMSE : 6.396890687830893


Training MAPE : 6.203861698916588
Testing MAPE : 6.22315646629609
```



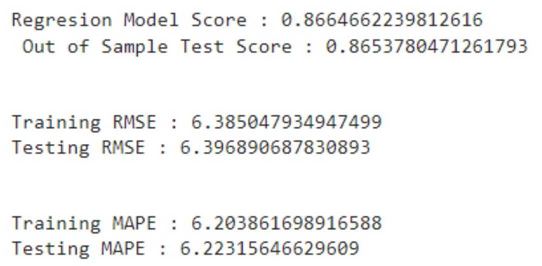
We can see that when we fitted the model in LASSO regression we came up 86% of both training and testing accuracy. RMSE and MAPE are also almost equal for both training and testing dataset.

Ridge Regression:

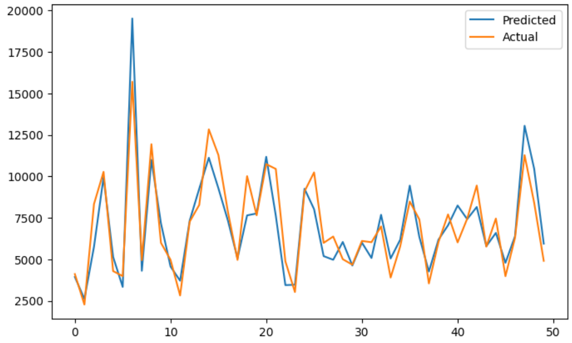
Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering. Also known as **Tikhonov regularization**, named for Andrey Tikhonov, it is a method of regularization of ill-posed problems. it is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters. In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias.

Model Performance:

We tried implementing our model in Ridge regression algorithm we came up with following outcomes.



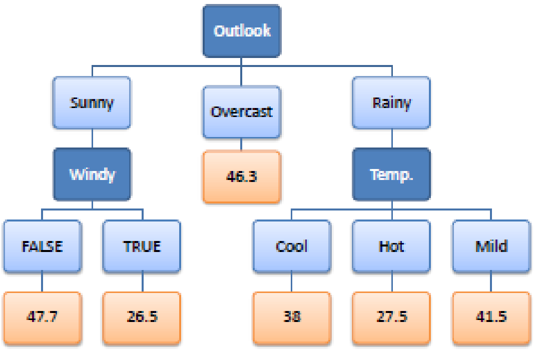
```
Regresion Model Score : 0.8664662239812616
 Out of Sample Test Score : 0.8653780471261793


Training RMSE : 6.385047934947499
Testing RMSE : 6.396890687830893


Training MAPE : 6.203861698916588
Testing MAPE : 6.22315646629609
```



We can see that when we fitted the model in Ridge regression we came up 86% of both training and testing accuracy. RMSE and MAPE are also almost equal for both training and testing dataset.

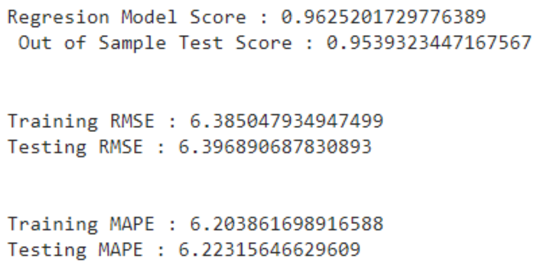
Decision Tree:

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

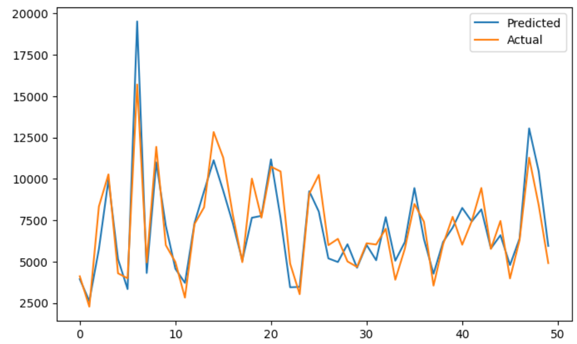


Model Performance:

We tried implementing our model in Decision Tree regressor algorithm we came up with following outcomes:



```
Regresion Model Score : 0.9625201729776389
 Out of Sample Test Score : 0.9539323447167567


Training RMSE : 6.385047934947499
Testing RMSE : 6.396890687830893


Training MAPE : 6.203861698916588
Testing MAPE : 6.22315646629609
```



Here we have the best model accuracy till now with 96% training and 95% testing accuracy. RMSE and MAPE also suggests good performance of the model.

Random Forest:

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

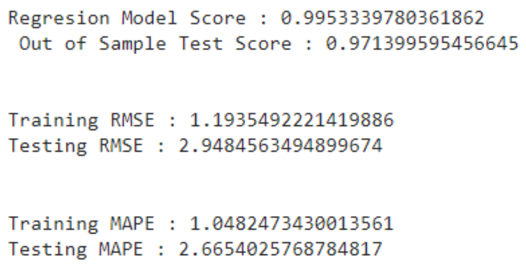
We need to approach the Random Forest regression technique like any other machine learning technique

- Design a specific question or data and get the source to determine the required data.
- Make sure the data is in an accessible format else convert it to the required format.
- Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
- Create a machine learning model
- Set the baseline model that you want to achieve
- Train the data machine learning model.
- Provide an insight into the model with test data
- Now compare the performance metrics of both the test data and the predicted data from the model.
- If it doesn’t satisfy your expectations, you can try improving your model accordingly or dating your data, or using another data modeling technique.

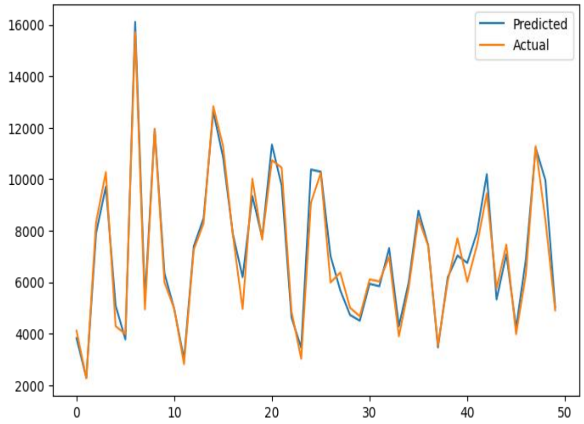
- At this stage, you interpret the data you have gained and report accordingly.

Model Performance:

We tried implementing our model in Random Forest bagging algorithm we came up with following outcomes**:**



```
Regresion Model Score : 0.9953339780361862
 Out of Sample Test Score : 0.971399595456645


Training RMSE : 1.1935492221419886
Testing RMSE : 2.9484563494899674


Training MAPE : 1.0482473430013561
Testing MAPE : 2.6654025768784817
```



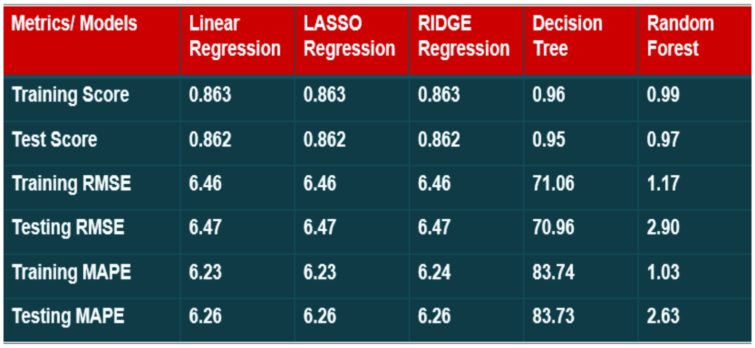
It is clearly evident that Random Forest surpasses even decision trees in terms accuracy. With 99% training and 97% testing accuracy this model performs best out of all the models we tried.

Conclusion:

The main objective of sales forecasting is to paint an accurate picture of expected sales. Sales teams aim to either hit their expected target or exceed it.

When the sales forecast is accurate, operations go smoothly and future planning for the company's growth is done efficiently.

Upon having this analysis it can be established that given the dataset, the model developed is able to explain 95.5878 % of the variations and is able to predict the sales values in a good range.



Metrics/ Models	Linear Regression	LASSO Regression	RIDGE Regression	Decision Tree	Random Forest
Training Score	0.863	0.863	0.863	0.96	0.99
Test Score	0.862	0.862	0.862	0.95	0.97
Training RMSE	6.46	6.46	6.46	71.06	1.17
Testing RMSE	6.47	6.47	6.47	70.96	2.90
Training MAPE	6.23	6.23	6.24	83.74	1.03
Testing MAPE	6.26	6.26	6.26	83.73	2.63

Random Forest has proven to be the most efficient model among the algorithms used in our model, including Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, and Random Forest, with 99% Training Accuracy and 97% Testing Accuracy.

Challenges:

The major challenge would be the computational time and RAM needed towork upon such a dataset in a cloud environment.

Recommendations:

On the basis of our analysis and prediction we came up with some recommendations which may work in the favor of the company to grow their business:

- According to a pattern we have noticed, stores that are far from their rivals are less likely to see sales than those that are close to one another. Therefore, we recommend that stores be opened close to those of its rivals.
- We discovered that the month of October has the highest sales across all three years, so for next year, some additional offers and promotions should be provided for that period to capitalize on the advantage.
- We discovered that customers in Assortment B spend a lot of money, but there aren't many of them. If some targeted ads are pushed to the wealthy customers, sales in this assortment can increase.
- Even if there is no promotion running over the weekend, there are still significant sales. It can be recommended to offer various incentives in the weekend to boost sales.
- Only Store Type B carries all three of the offered assortments (a, b, and c). Stores for c type assortment levels are lacking for other store types, so efforts should be made to fill these gaps in order to increase revenue.

Acknowledgement:

This project is presented by-**Kumar Abhinav** and **Saumya Dash**. Our sincere efforts have made us to accomplish the task of completing this project. We are extremely grateful to our instructors and mentors who have helped us to grow in this field. We would like to express our sincere gratitude to the celebrated authors whose phenomenal work has been consulted and referred in our project work. We also wish to convey our appreciation to our peers who provided encouragement and timely support in the hour of need. This project helped

us improve our skills and enhanced our knowledge.

Reference:

- Machine Learning Mastery
- GeeksforGeeks
- Analytics Vidhya Blogs
- Towards Data Science Blogs
- Built in Data Science Blogs
- Scikit- Learn Org
- Investopedia