Retail Sales Prediction

**Abstract:**

Forecast prediction is a process of predicting a future value using past data and some other related factors. If we want to forecast the present day sales or future sales using past sales and some other related factors like seasonality, festivals, economic conditions etc. then it is known as Sales forecasting.

We can use different methods for Sales forecasting. Among them, one of the most robust methods is Sales forecasting using machine learning algorithms.

In this method, we preprocess all the input to machine-understandable features and then apply a machine learning algorithm on top of it.

**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:**

Demand for product or service is not constant, it changes with respect to time. But to maintain the demand and supply balance it is very important to understand the demand of product or service in the future.

Sales prediction is a process of estimating demand of sales of a particular product or service over a period of time.

Sales prediction not only helps in balancing the supply chain but also helps in making future business strategies like budgets, hiring, incentives, goals, acquisitions and various other growth plan.

**Overview of data:**

There are two datasets

#### Rossmann Stores Data.csv - historical data including Sales

#### store.csv - supplemental information about the stores

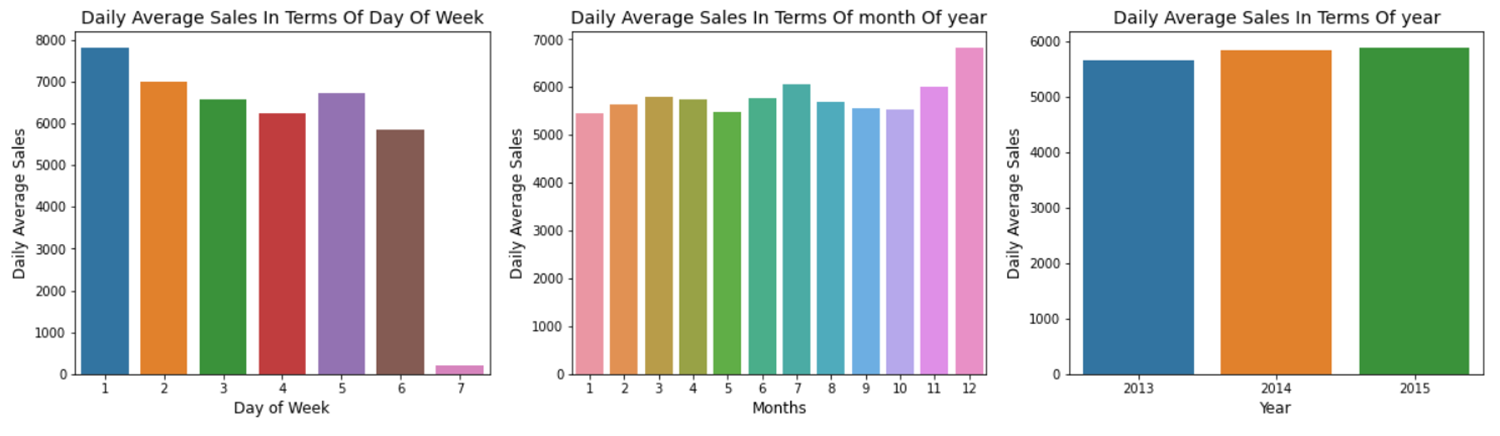
The Features of Dataset:

* **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

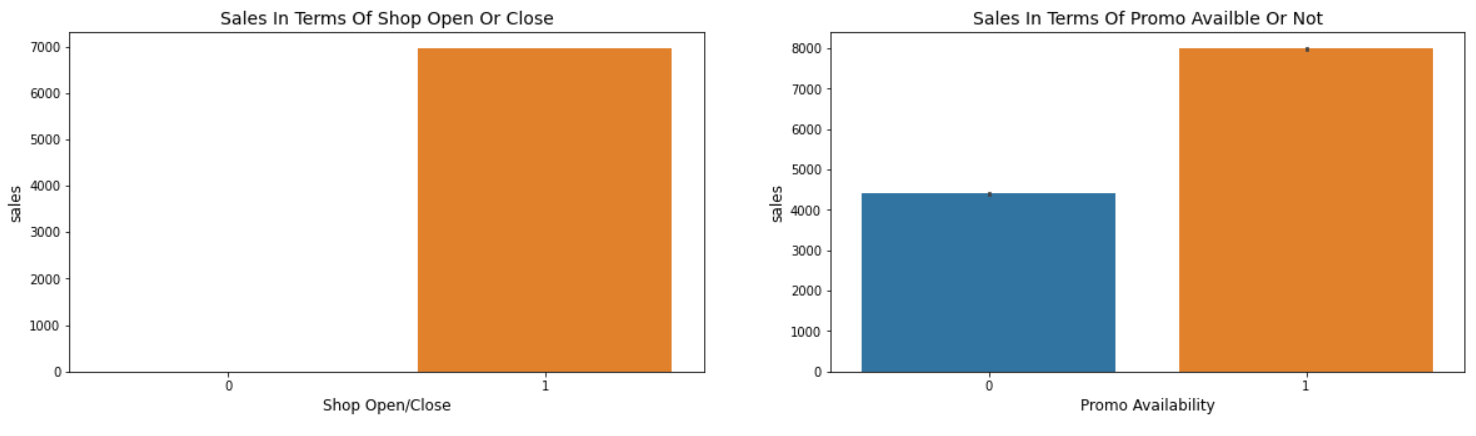
**Steps Involved**

**1.EDA (Exploratory data analysis):**

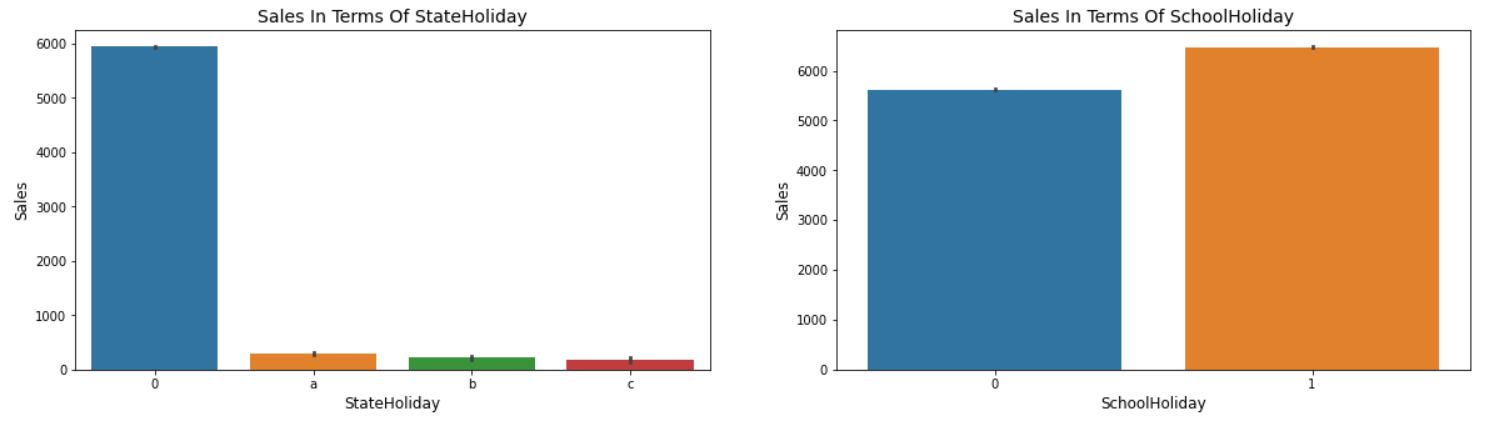
1. Analysing the basic information of the dataset like number of observations and features, data Type of different features, null values of each feature.
2. Merging both the dataset and understanding the important insights of the data.



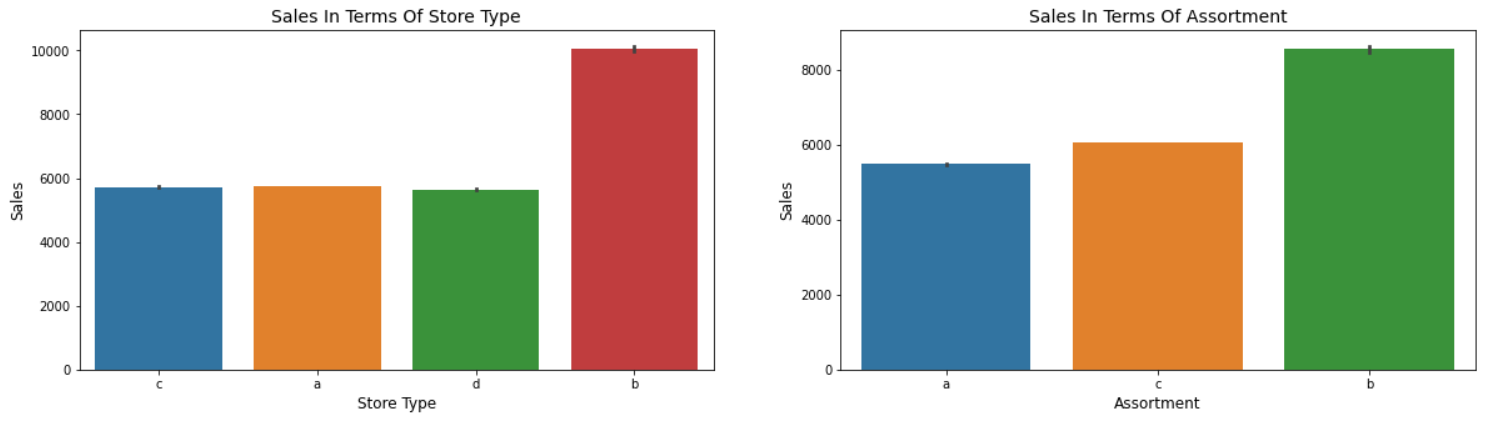
* First represents daily average sales in terms of day of week. Values 1-7 represent each day of the week. On day 7 the sales are very low because most of the stores are closed as it is Sunday. Sales are high on Monday, most probably because many stores are closed on Sunday.
* Second represents daily average sales in terms of month of year. Month 12 that is December has the highest sales.
* Third represents daily average sales in terms of year.



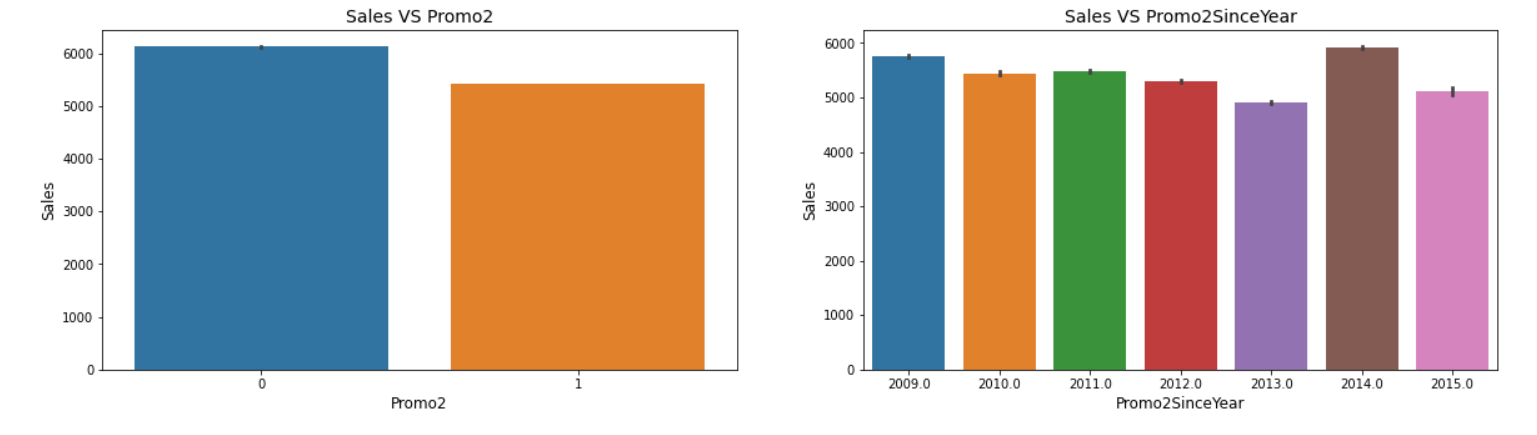
* First graph shows sales in terms of shop open or close. 0 represents shop is basically closed so there is no sale on that day.
* Second graph shows sales in terms of promo available or not. Sales are pretty high when promo is available.

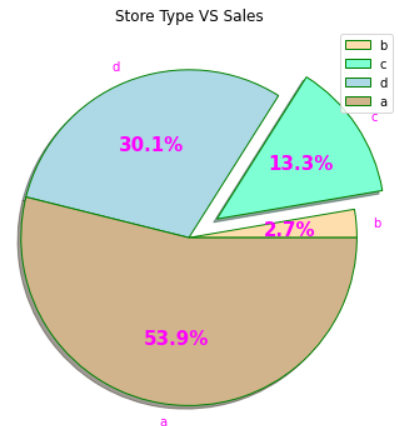


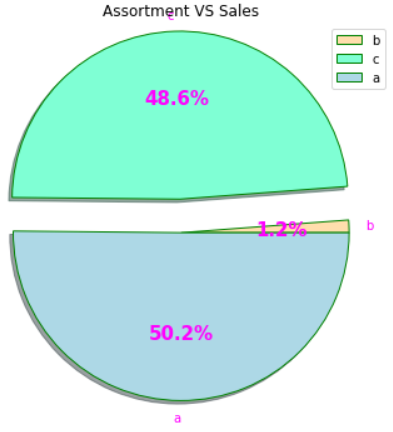
* First graph indicates sales in terms of state holiday. Normally all stores, with few exceptions, are closed on state holidays. a = public holiday, b = Easter holiday, c = Christmas, 0 = None. Lowest of the sales were seen on state holiday especially on Christmas.
* Second graph indicates sales in terms of school holiday. More stores were open on School Holidays than on State Holidays and hence had more sales than State Holidays.

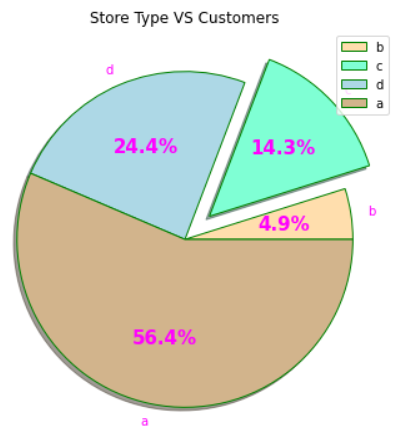


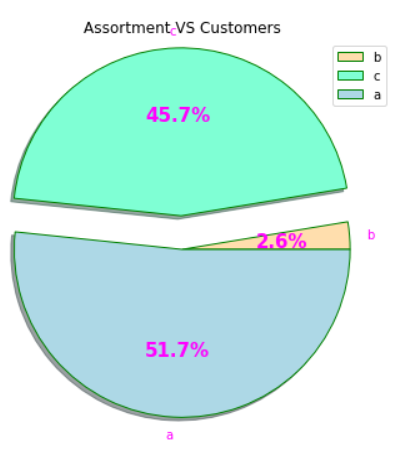
* Graph 1 - On an average store type B had the highest sales. There must be something different about this store type.
* Graph 2 - Highest average sales were seen with Assortment levels-b which is 'extra'.

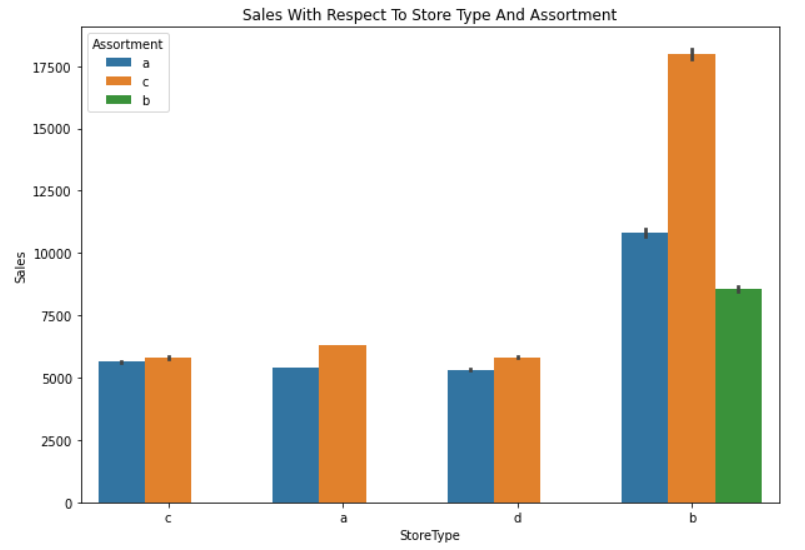


* Graph 1 - With Promo2, slightly more sales were seen without it which indicates there are many stores not participating in promo.
* Graph 2 - Describes the year when the store started participating in Promo2.

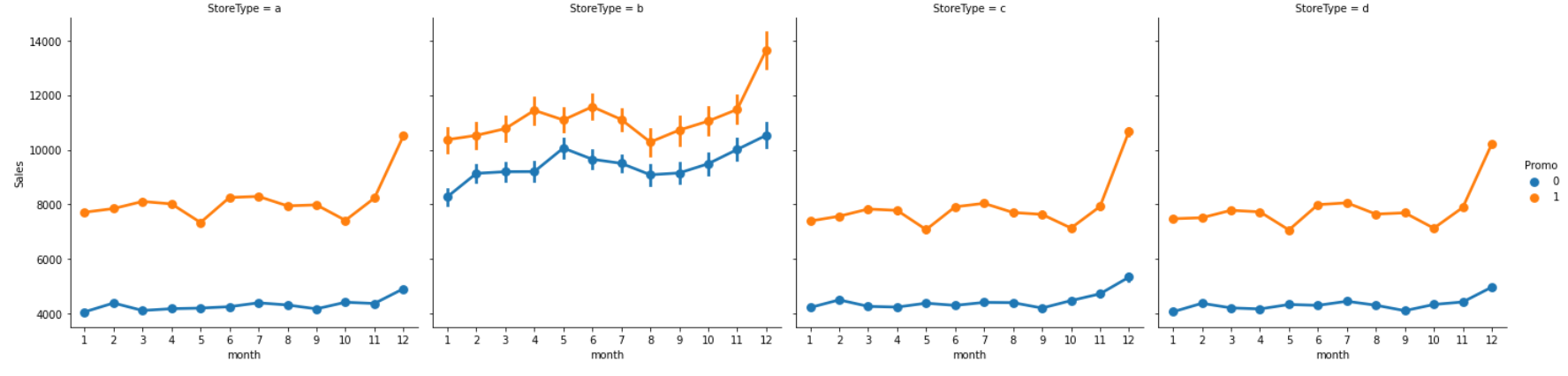




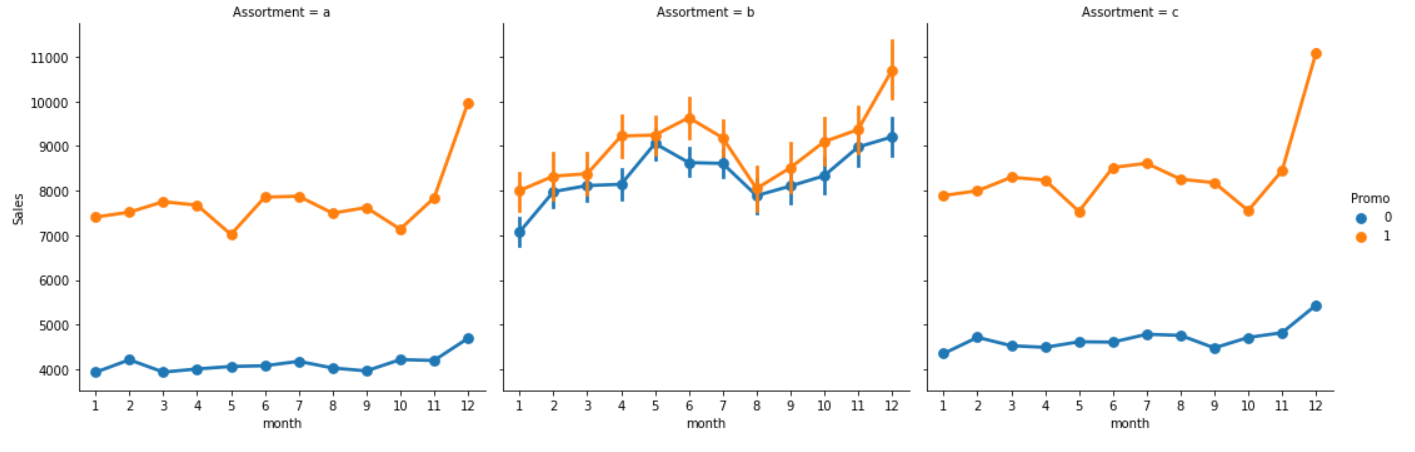


* A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. Earlier it was seen that the store type 'b' had the highest sales on an average because the default estimation function to the bar plot is mean.
* But upon further exploration it can be clearly observed that the highest sales belonged to the store type 'a' due to the high number of type 'a' stores in our dataset. Store type 'a' and 'c' had a similar kind of sales and customer share.

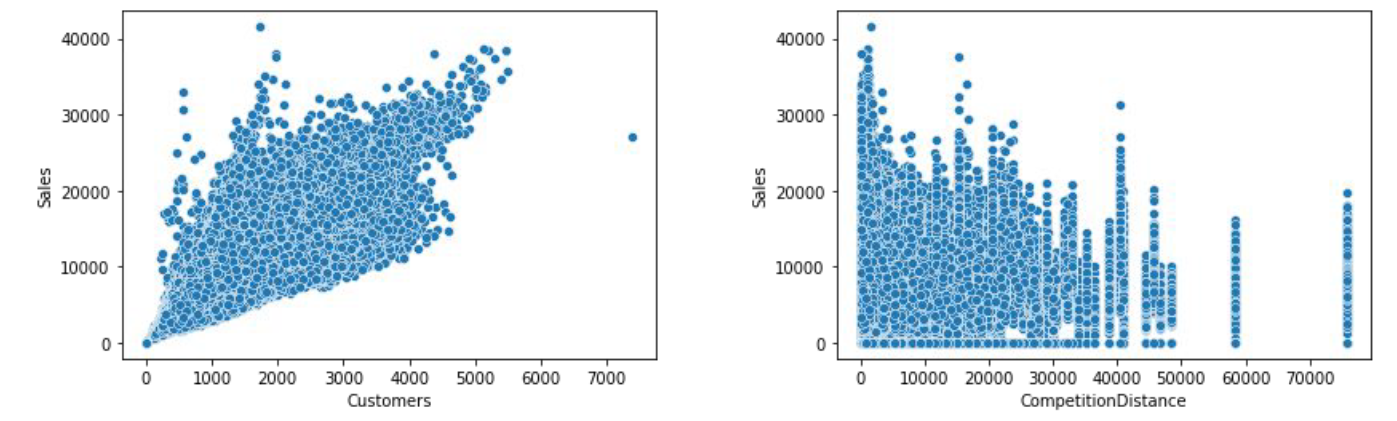
Interesting insight to note is that store type b with highest average sales and per store revenue generation looks healthy and a reason for that would be all three kinds of assortment strategies involved. On the other hand store type a, c and d have only assortment level a and c.



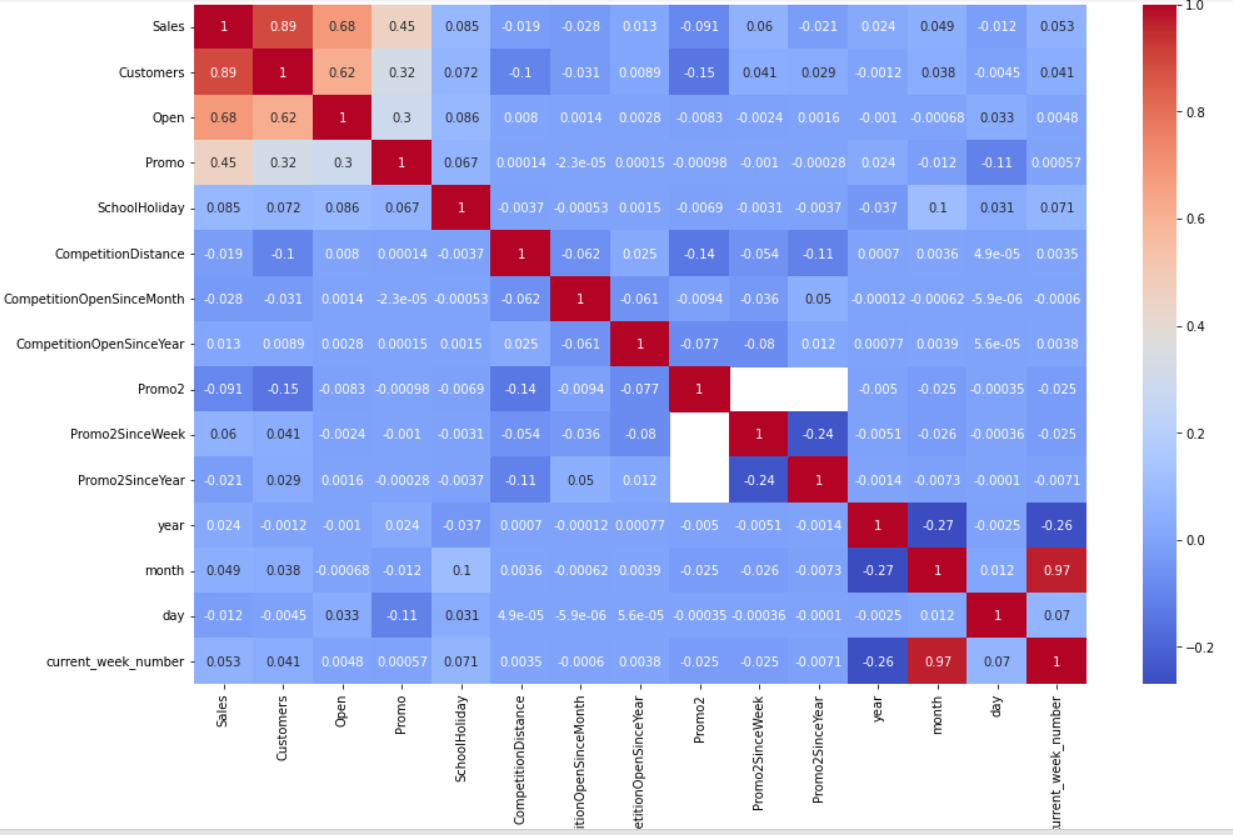
Monthly Sales in terms of Store type and Promo. When there is promo, sales of all the store type have increased.



Monthly Sales in terms of Promo and Assortment. When there is promo, sales of all the assortment have increased.



* Pair plot of Sales v/s Customer shows that sales increase as the number of customers increase.
* Pair plot of Sales v/s Competition distance tells that sales increase when the competition stores are nearer.



* Day of the week has a negative correlation indicating low sales as the weekends, and promo, customers and open has positive correlation.
* State Holiday has a negative correlation suggesting that stores are mostly closed on state holidays indicating low sales.
* 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.

**2. Data Manipulation and Feature Engineering**

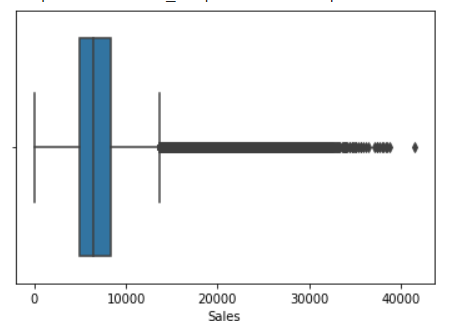
* Data manipulation involves manipulating and making necessary changes in our dataset before deploying it into machine learning algorithm.
* Extracting current date, month, year, week number from date column.
* Since there is no sale when the shops are closed, we removed all the observations when the store is closed.
* Combine CompetitionOpenSinceMonth, CompetitionOpenSinceYear to give "competition\_open” which tells since how many months competition is open.
* Combine Promo2SinceWeek, Promo2SinceYear to give "promo\_2\_open“which tells since how many months the shop is participating in promo2.
* Getting “IsPromo2Month” from promo\_interval\_open which tells is Promo2 open for a particular month or not.
* CompetitionDistance has some null values we will deal with it by filling null values with median of CompetitionDistance

**Outlier Detection:**

An unusual high or low values of data as compared to other observations can be called as outliers. In this project we have used 2 methods to detect outliers one is using IQR (Interquantile range) method. Second is using zscore.

Using IQR

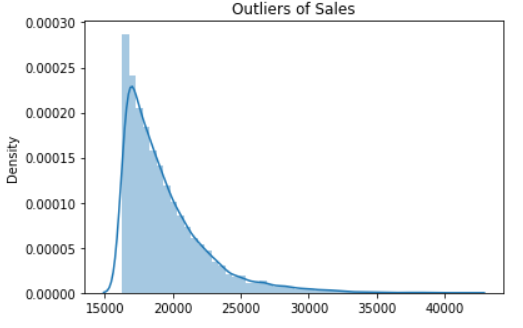
* Data points that fall below the lower bond and data points that fall above the upper bond are considered as outliers. we calculate it using IQR interquantile range.



* Getting 25th and 75th percentile of values which can be called as quantile1 and quantile 3 respectively.
* IQR value=quantile3-quantile1
* Lower bond=quantile1-1.5\* IQR value
* Upper bond=quantile3+1.5\* IQR value
* Finally plotting box plot to detect outliers.

Using zscore

* Z-score tells how many standard deviations away a given observation is from the mean.
* z = (x-mean)/standard deviation
* More than 3 standard deviations are considered as an outlier.

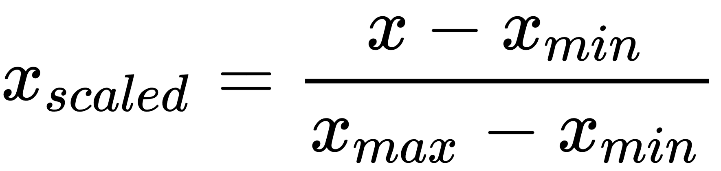


**Feature Scaling:**

Feature Scaling is a technique to standardize the independent features present in the data within a fixed range. It is done to prevent biased nature of machine learning algorithms towards any particular feature. There are two techniques to do this

* 1. Normalization
  2. Standardization

The technique which we have used is Normalization or Min-Max Scaling. It helps to rescale values in the range of 0 & 1. The equation to do so is



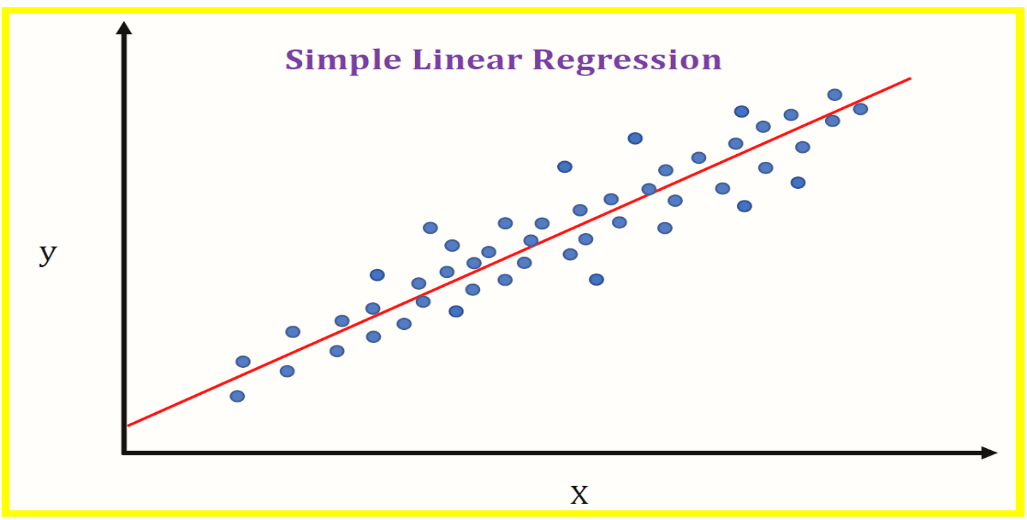
**One hot Encoding:**

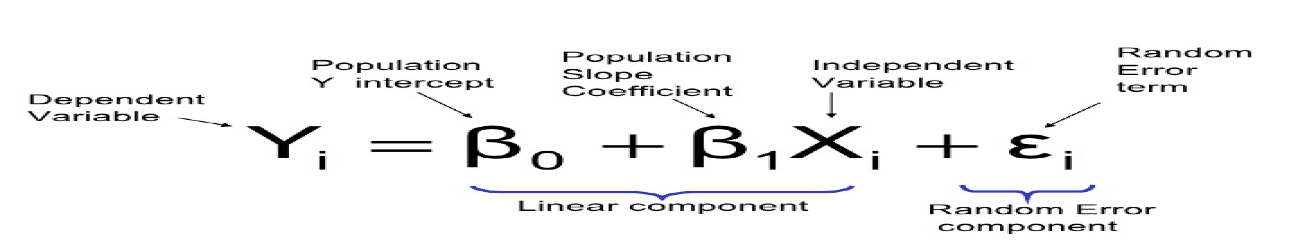
For categorical variables we need to perform one hot encoding to convert it into numeric so that it can be deployed into machine learning algorithm.

**Algorithms Used:**

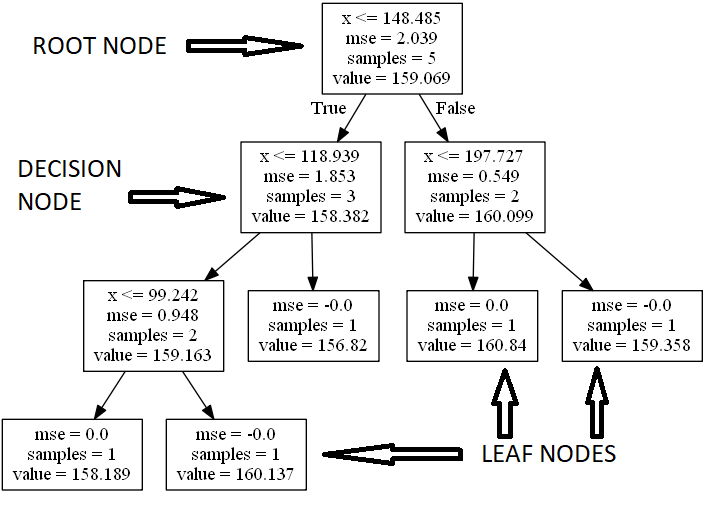
1. **Linear Regression**

Linear Regression is a machine learning algorithm grounded on supervised learning. It performs a regression task. Regression models a target prediction value grounded on independent variables. It's substantially used for finding out the relationship between variables and prediction. Different regression models differ grounded on – the kind of relationship between dependent and independent variables they're considering, and the number of independent variables getting used.



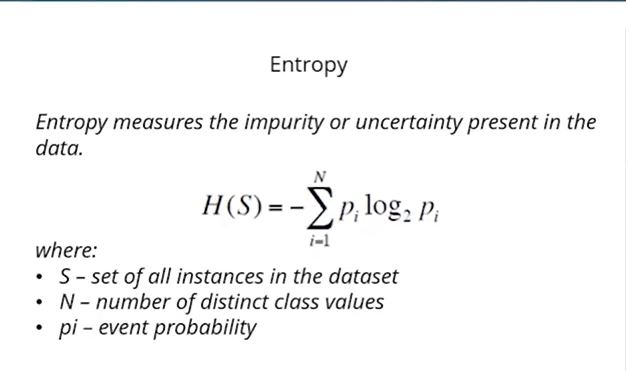
Equation for liner equation is given by

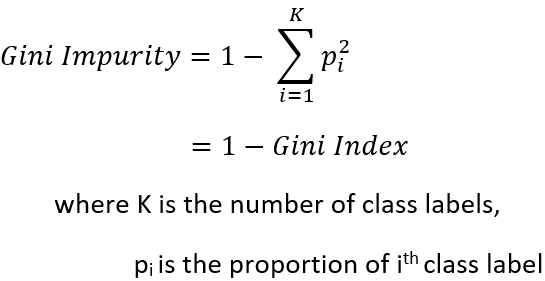
1. **Decision Tree**

It is the most impressive and well-known ML algorithm for classification and prediction. A Decision tree is a flowchart-like tree structure, where each inner node signifies a test on an attribute, each branch addresses a result of the test, and each leaf node (terminal node) holds a class label.

The model used a different method for accurate splitting of nodes-

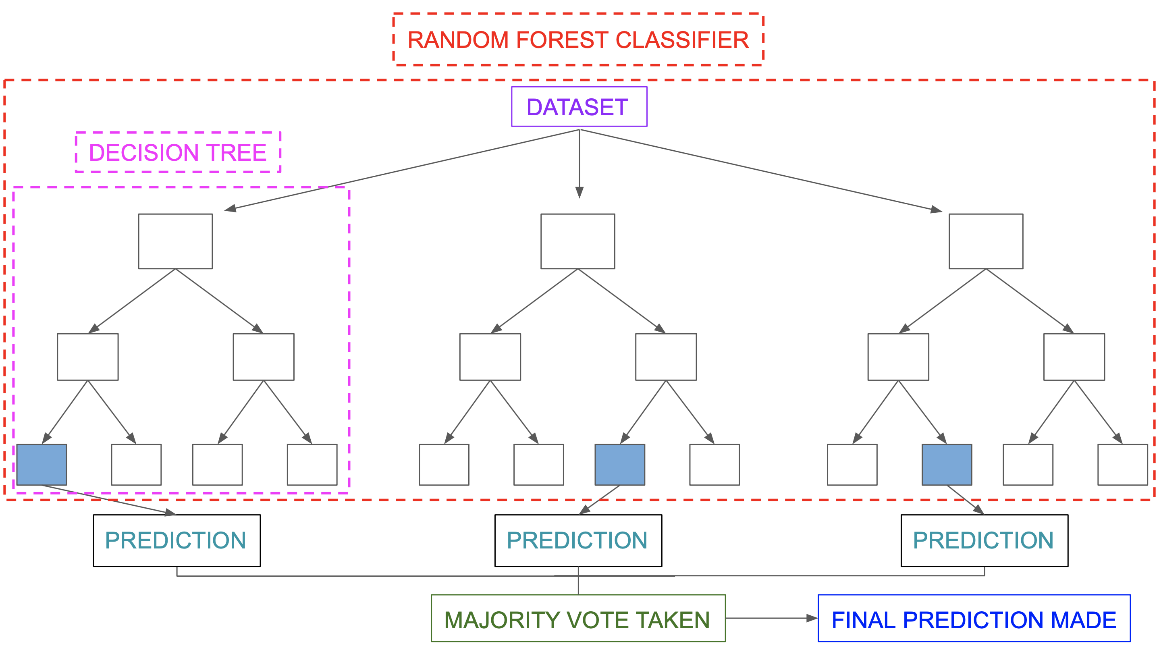
1. Entropy-Entropy can be characterized as a proportion of the purity of the sub-split. Entropy generally lies between 0 to 1. Entropy can be calculated using the formula -



2. Gini Impurity: The inward working of Gini impurity is additionally fairly like the working of entropy in the Decision Tree. The Gini Impurity of highlights subsequent to parting can be determined by utilizing this formula.

1. **Random Forest**

Every decision tree has an excessive variance, however, while we integrate them all collectively in parallel then the ensuing variance is low as every selection tree receives perfectly trained on that sample data, and as a result, the output doesn’t rely upon one decision tree however on more than one decision trees. In the case of a classification problem, the final output is taken through the use of majority voting. In the case of a regression problem, the final output is the average of output of different tress.



Random Forest is an ensemble approach able to act each regression and classification with the usage of multiple discussion trees and a method referred to as Bootstrap and Aggregation, generally referred to as bagging.

1. **Gradient Boosting**

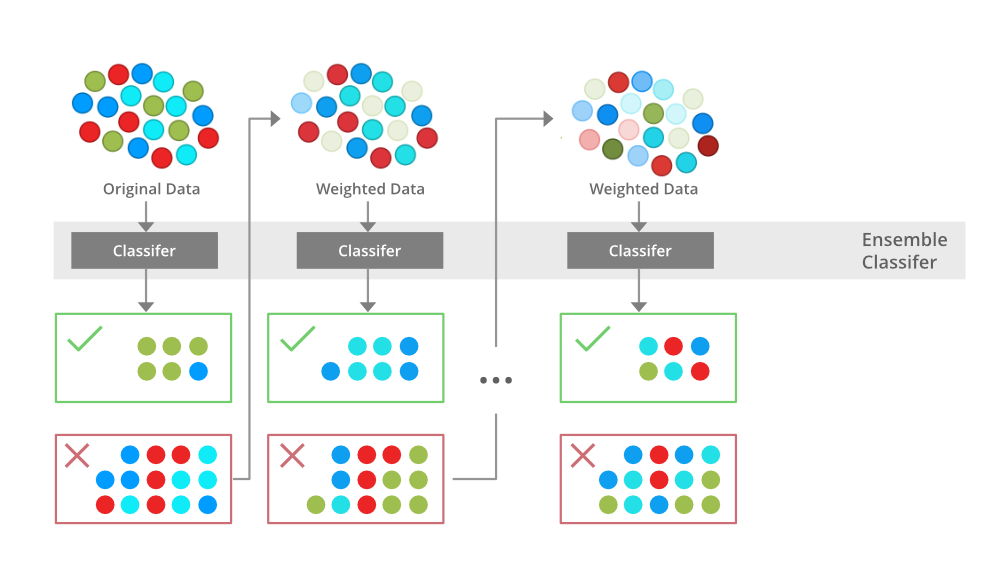
Gradient Boosting is an ensemble modelling, approach that tries to construct a robust classifier from the range of weak classifiers.

Gradient Boosting is a famous boosting algorithm. In gradient boosting, every predictor corrects its previous error.

It also supports parallelization (it can generate the different nodes of tree parallel), can

use the hardware resources efficiently through Cache-Awareness, it can also handle

sparse data efficiently.



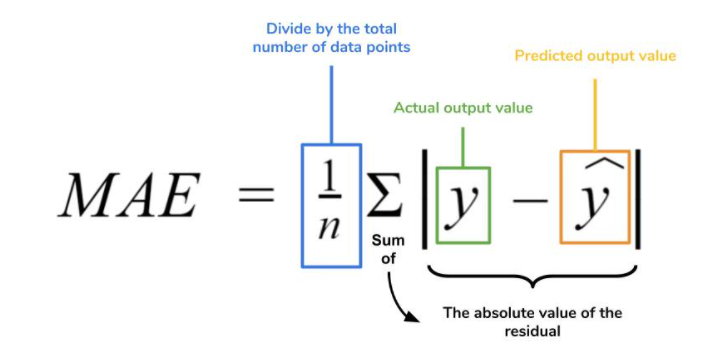
**Model Performance:**

After the model is built, if we see that the difference in the values of the predicted and actual data is not much, it is considered to be a good model and can be used to make future predictions.

Few metric tools we can use to calculate error in the model

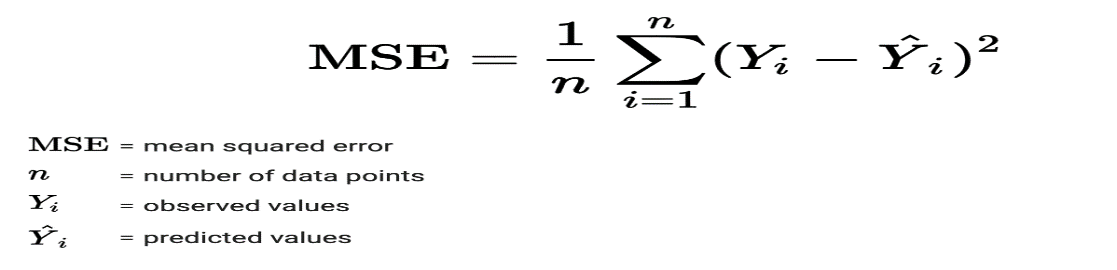
1. **MAE (Mean Absolute Error)**

MAE is the absolute difference between the actual value and the value predicted by the model. The MAE is more robust to outliers as it takes absolute difference and does not penalize the errors as that of MSE. MAE is a linear score which means all the individual differences are weighted equally.



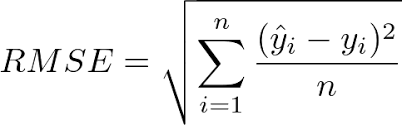
1. **MSE (Mean Squared Error)**

MSE simply takes average of the squared difference between the actual value and the value predicted by the regression model.



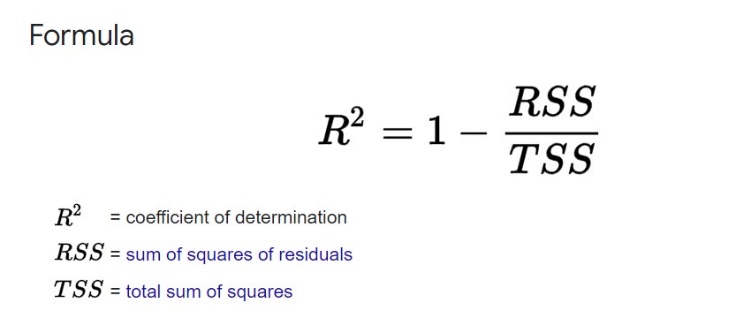
1. **RMSE (Root Mean Squared Error)**

RMSE is a metric which takes square root of the averaged squared difference between the target value and the value predicted by the model. It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors.



1. **R^2 (R - Squared)**

R² also known as Coefficient of Determination is another metric used for evaluating the performance of a regression model. The metric helps us to compare our current model with a constant baseline and tells us how much our model is better. The constant baseline is chosen by taking the mean of the data and drawing a line at the mean. R² will always be less than or equal to 1.



**Hyper Parameter Tuning:**

A machine learning model is defined as a mathematical model with a number of parameters to be learned from data. By training the model with existing data, we are able to adjust the model parameters.

However, there is another kind of parameters, known as hyperparameters, which cannot be directly learned from the normal training process. They are usually fixed before starting the actual training process.

Models can have many hyperparameters and finding the best combination of parameters is important. The two best strategies for tuning hyperparameters are:

1. **Grid Search CV**

In GridSearchCV approach, the machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for the best set of hyperparameters from a grid of hyperparameters values. The major drawback of GridSearchCV is that it will go through all the combinations of hyperparameters which makes grid search computationally very expensive.

1. **Randomized Search CV**

RandomizedSearchCV0 solves the disadvantage of GridSearchCV, as it goes through only a fixed number of hyperparameters. It moves within the grid in a randomly to find the best set of hyperparameters. This approach reduces unnecessary computation. The disadvantage of this method is the combinations the hyperparameter choose is beyond our control.

**Conclusion:**

Sales Prediction helps in making future business strategies like budgets, hiring, incentives, goals, acquisitions and various other growth plan. In this project we analyzed more than one thousand stores for sales prediction. After analysing we conclude some important observations as follows

1. Stores which are running promo have more sales.
2. The State Holiday affects adversely to sales while school holiday affects positively to sales.
3. Store type B though being few in number had the highest sales average. The reasons

include all three kinds of assortments specially assortment level b which is only

available at type b stores and being open on Sundays as well.

1. With increase in competition distance sales decrease. This may be because the store with low competition distance indicates that the store is in busy place.

**References:**

1. Machine Learning Mastery
2. GeeksforGeeks
3. Analytics Vidhya Blogs
4. Towards Data Science Blogs