Retail Sales Prediction

# By Mayank Ghai

# Data Science Trainee AlmaBetter, Bangalore

**Abstract:**

Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period

Businesses use sales forecasts to determine what revenue they will be generating in a particular timespan to empower themselves with powerful and strategic business plans. Important decisions such as budgets, hiring, incentives, goals, acquisitions, and various other growth plans are affected by the revenue the company is going to make in the coming months, and for these plans to be as effective as they are planned to be it is important for these forecasts to also be as good.

The sales forecasts are also different from the sales goals a company has. Sales-goals is are what a company wants to happen to execute its plans for the business. On the other hand, sales forecasts are what is going to happen based on past re, data, trends, and various improvement measures taken.

The work here predicts the sales for a drug store chain in the European market for six weeks and compares the results of machine learning algorithms.

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

# Problem Statement:

Rossmann operates over 3,000 drug stores in 7

European countries. Rossmann store managers are currently tasked with predicting their daily sales 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.

# Introduction:

The interest in a product continues to change occasionally. No business can work on its monetary growth without precisely assessing client interest and future demand for items.

Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period.

For a good sales forecast, it is extremely important to get a good dataset as well. Forecasts heavily depend on the records, trends, and patterns observed for sales of a particular store. The variations could be due to several reasons.

Talking from a business’s point of view, these sales forecasts are done consistently to improve

their sales forecasting models as they directly impact their decision-making process, goals, plans, and growth strategies.

In this Retail Sales Prediction, machine learning models have been created that predict sales of these 1115 drug stores across the European market and compare the results of these models. In addition, an effort has been made to analyze and find all the features that are contributing to higher sales and the features that lead to lower sales, so that improvement plans can be worked upon.

# Approach:

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

## Understanding the business problem and 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 to work 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 of categorical features.
* **Modeling**- To know which model will be best for the prediction we have split the data into Test-Train data and with its help we checked the parameter that is used to choose the best model among the given predictive models.

## Model Performance and Evaluation

* **Store-wise Sales Predictions**
* **Conclusion and Recommendations**

# Understanding the Data:

The first step involved understanding the data and getting answers to some basic questions like; What is the data about? How many rows or observations are there in it? How many features are there in it? What are the data types? Are there any missing values? And anything that could be relevant and useful to our investigation. Let’s just understand the dataset first and the terms involved before proceeding further.

Our dataset consists of two CSV files, the first consists of historical data with 1017209 rows or observations and 9 columns with no null values. The second dataset was supplementary information about the stores with 1115 rows and 10 columns and a lot of missing values in a few columns. The data types were integer, float, and object in nature.

Let’s define the features involved:

* **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
* **State Holiday -** 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
* **School Holiday -** indicates if the (Store, Date) was affected by the closure of public schools
* **Store Type -** 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.
* **Competition Distance** – the distance in meters to the nearest competitor store
* **Competition Open Since[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

* **Promo2 Since[Year/Week] -** describes the year and calendar week when the store started participating in Promo2
* **Promo Interval** - 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, and November of any given year for that store.

**Data Cleaning and Preprocessing:** 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.

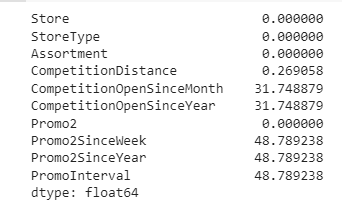
It is necessary to check and handle these values before feeding them to the models, to obtain good insights into what the data is trying to say and to make great characterization and predictions which will in turn help improve the business's growth.

The historical records dataset had no null values

.



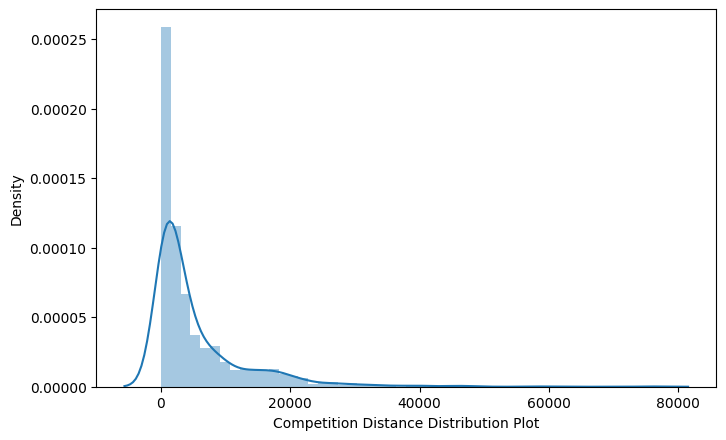
This below shows the percentage of the null data available in the store dataset.



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

* Competition Open Since Month
* Competition Open Since Year
* Promo2 Since Week
* Promo2 Since Year
* Promo Interval
* Competition Distance’ - 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



It seems like most of the values of the Competition Distance are towards the left and the distribution is skewed on the right. Median is more robust to outlier effect hence median was imputed in the null values.

Right skewed distributions occur when the long tail is on the right side of the distribution also called a positively skewed distribution which essentially suggests that there are positive outliers far along which influences the mean.

It seems like most of the values of the Competition Distance in the column are between 0-10kms. Consequently, the long tail in an asymmetrical distribution pulls the mean away from the most common values. The mean is greater than the median. The mean overestimates the most common values in the distribution and hence median is used in this case, it is more robust to the outlier effect and hence median is used to impute the missing values in this feature.

* CompetitionOpenSinceMonth- gives the approximate 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 the column is used to impute the missing values in the column as it gives the most occurring 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 with 0.

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

# Data Manipulation:

Data manipulation involves manipulating and changing our dataset before feeding it to various regression machine learning models. This involves keeping important features, outlier treatment, feature scaling, and creating dummy variables if necessary.

## Feature Engineering:

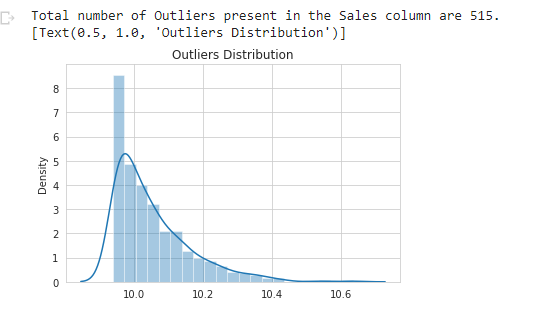
* Some stores were closed due to refurbishment and some on account of week off or holidays. Those stores on those dates generated zero sales and hence removing the rows was important to avoid confusion by the algorithms and then removing the feature altogether because it wasn’t providing any value in the prediction of the sales.
* There were features that like Competition Open since Month and Year. It was combined to count the total months since the nearest competition was opened.
* Promo2SinceWeek, Promo2SinceYear indicated promotion 2 opened since week and year. These features were combined to count the total months since promotion 2 is run.
* PromoInterval indicated the months for promotion 2 renewal. Hence, the sale month was compared against the interval and a new feature was created to determine whether the promo2 was renewed in that month.

## Outlier Detection:

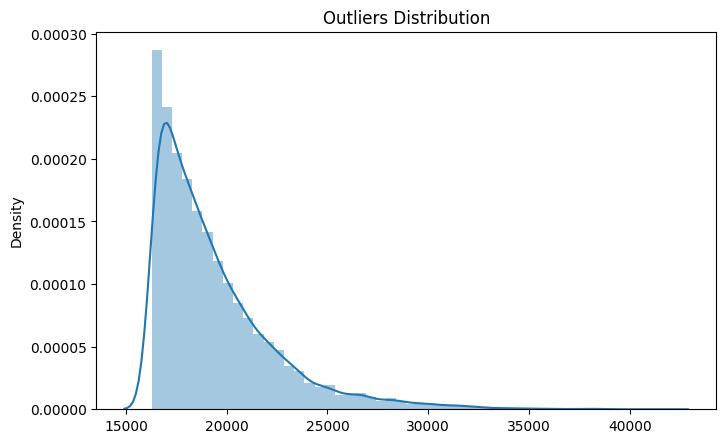
In statistics, an outlier is a data point that differs significantly from other observations. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.

Z-score is a statistical measure that tells you how far a data point is from the rest of the dataset. In a more technical term, Z-score tells how many standard deviations away a given observation is from the mean.

z = (x-mean)/standard deviation



**The data points with sales values higher than 10.2 are very low and hence they can be considered outliers. The percentage of outliers in our dataset**.



More than 3 standard deviations were considered an outlier. Exploring the outliers data frame, some important insights were generated:

* The data points with sales value higher than 28000 are very low and hence they can be considered outliers.
* The outliers had a day of the week as 7 i.e. Sunday and the store type for that observation was ‘b’.
* Other outliers had promotions running on that day.
* It can be well established that the outliers are showing this behavior for the stores with promotion = 1 and store type

B. It would not be wise to treat them because the reasons behind this behavior seem fair.

* Being open 24\*7 along with all kinds of assortments available is probably the reason why it had higher average sales than any other store type.
* If the outliers are a valid occurrence it would be wise not to treat them by deleting or manipulating them especially when we have established the ups and downs of the target variable about the other features. It is well established that there is seasonality involved and no linear relationship is possible to fit. For these kinds of datasets tree-based, machine learning algorithms are used which are robust to outlier effect.

# Exploratory Data Analysis:

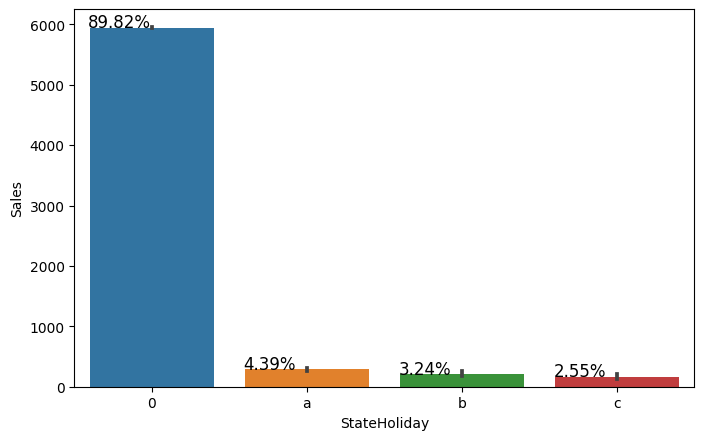
Exploratory data analysis is a crucial part of data analysis. It involves exploring and analyzing the dataset given to find out patterns, trends, and conclusions to make better decisions related to the data, often using statistical graphics and other data visualization tools to summarize the results. 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 factors might be contributing to the high and low sales numbers.

## Approach:

There are two kinds of features in the dataset: Categorical and Non-Categorical Variables.

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

Non-Categorical- A noncategorical or continuous variable is a variable whose value is obtained by measuring, i.e., one which can take on an uncountable set of values.

Both of them are analyzed separately. Categorical data is usually analyzed through count plots and bar plots by the target variable and that is what is done here too. On the other hand, Numeric or Continuous variables were analyzed through distribution plots, box plots, and scatterplots to get useful insights.

## Hypotheses

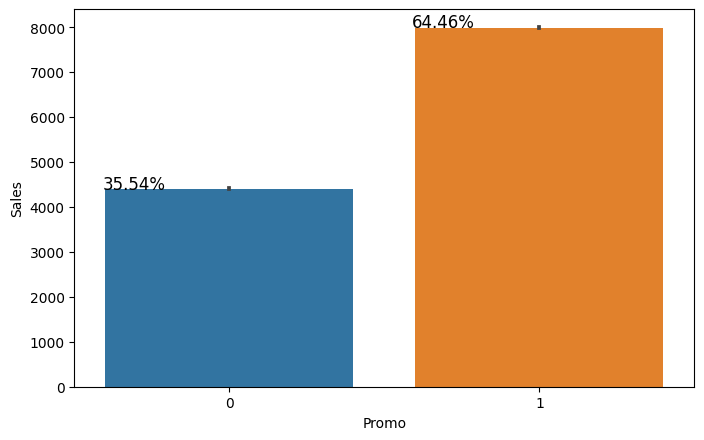
Just by observing the head of the dataset and understanding the features involved in it, the following hypotheses could be framed:

* There's a feature called "DayOfWeek" with the values 1-7 denoting each day of the week. There would be a week off probably Sunday when the stores would be closed and we would get low overall sales.
* Customers would have a positive correlation with Sales.
* The Store type and Assortment strategy involved would be having a certain effect on sales as well. Some premium high-quality products would fetch more revenue.
* Promotion should be having a positive correlation with Sales.
* Some stores were closed due to refurbishment, which would generate 0 revenue for that period.
* Stores are influenced by seasonality, probably before holidays sales would be high.

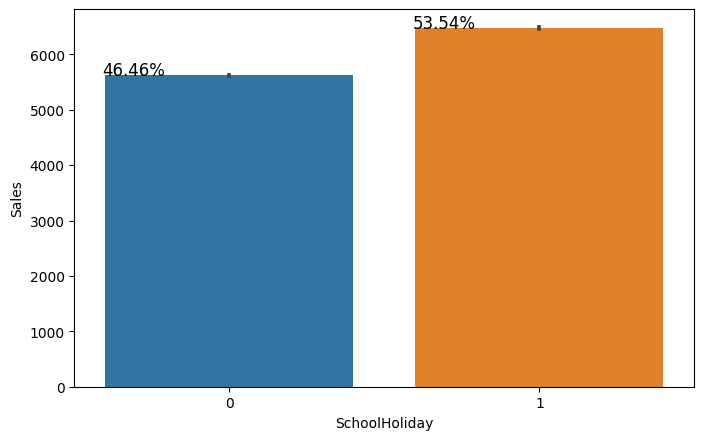
## Categorical Insights:

## 

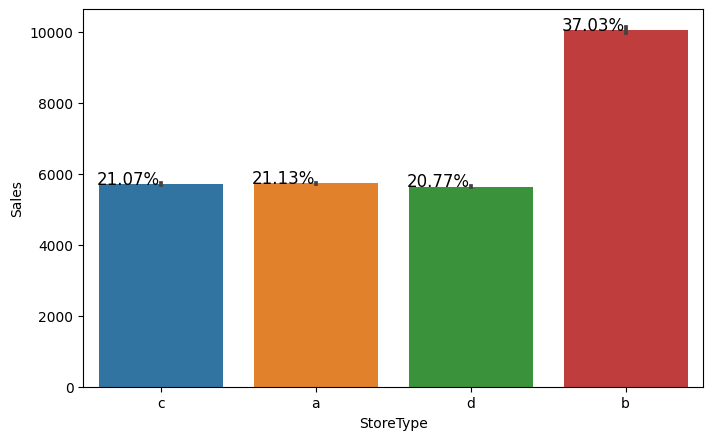
Here it can be deduced that there were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.

The promotion has a positive effect on Sales indicating high sales for stores with Promo.

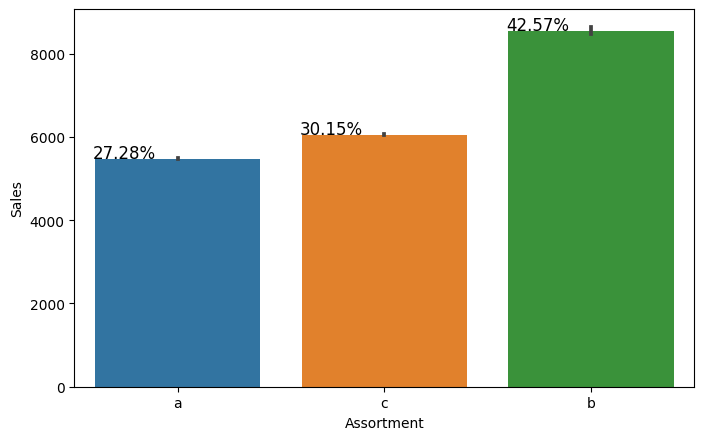
Sales were low whenever a State Holiday was indicating only a few stores were open on these days.



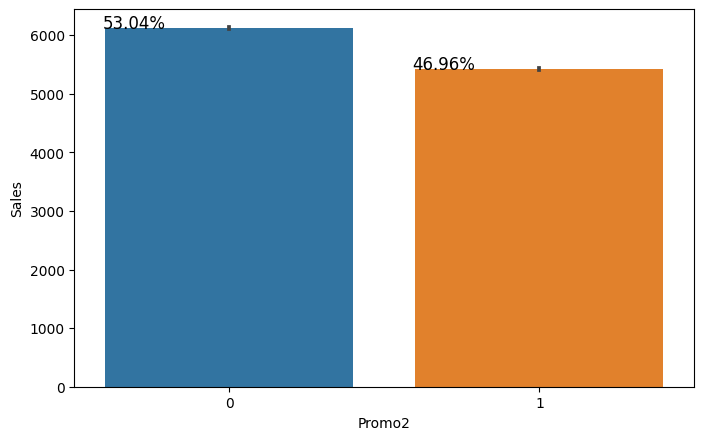
Sales were high on average on School Holidays indicating School Holidays weren’t compulsory by the law and comparatively more sales were recorded on holidays.



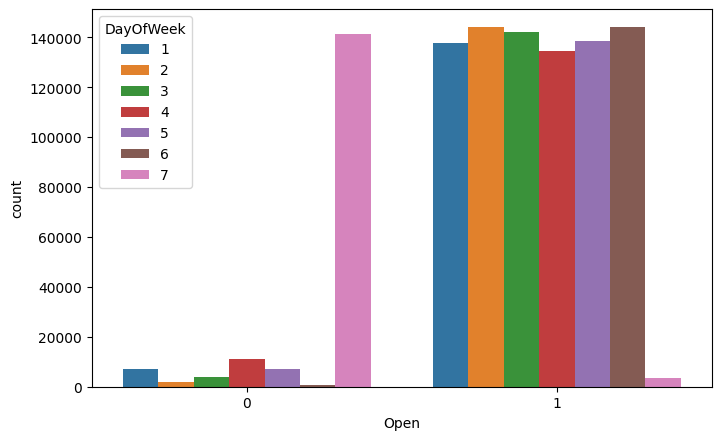
A bar plot represents an estimate of the central tendency for a numeric variable with the height of each rectangle. The store type b has the highest sales on average.



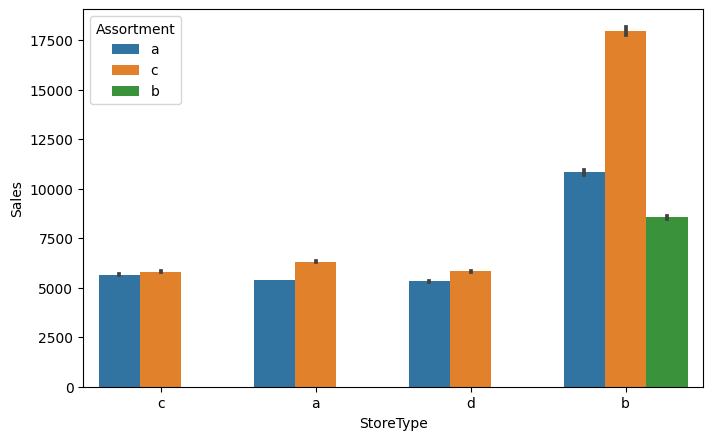
Assortment type b has the highest sales on average.



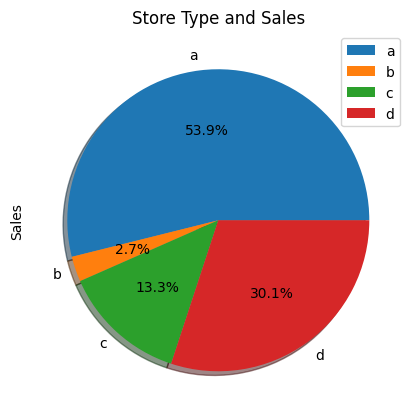
With Promo2, slightly more sales were seen without it which indicates there are many stores not participating in the promo.



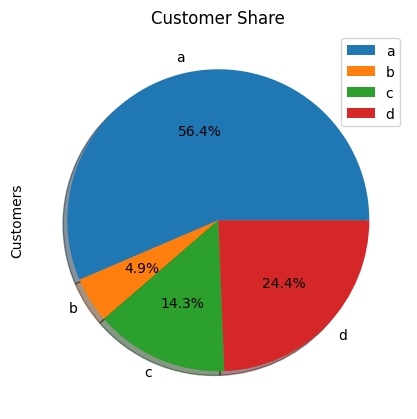
This is a count plot of open shops according to the day of the week. The number of shops open on Sundays was very less and hence low sales. Some shops were closed on weekdays as well accounting for the stores closed due to refurbishment or holidays.



The above bar plot shows that the store types a, c and d have only assortment levels a and c. On the other hand, store type b has all three kinds of assortment strategies, a reason why average sales were high for store type b stores.

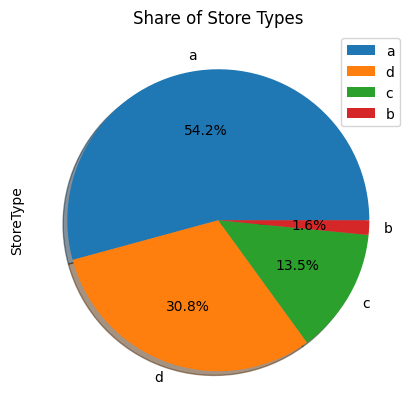


When plotting a pie chart for the sum of sales of the various store types, it can be observed that even though type a stores had the most sales, type b stores were high on average.

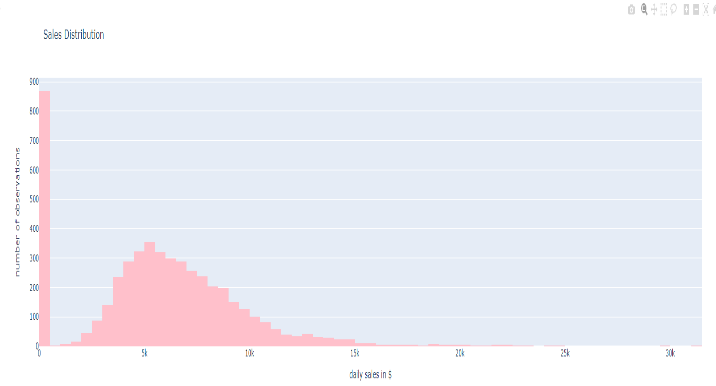


But upon further exploration, it can be observed that the highest sales belonged to the store type due to the high number of type a stores in our dataset. Store types **a** and **c** had a similar kind of sales and customer share.

An interesting insight to note is that store type **b** with the highest average sales and per store, revenue generation looks healthy and a reason for that would be all three kinds of assortment strategies involved which was seen earlier.

Based on the above findings it seems that there are quite a lot of opportunities in store type 'b' & 'd' as they had more number of customers per store and more sales per customer, respectively. Sttypestypes a & c are quite similar in terms of "per customer and per store" sales numbers and just because the majority of the stores were of these kinds, they had the best overall revenue numbers. On the other hand, store type b were very few and even then they had better average sales than other.

## Continuous Insights:

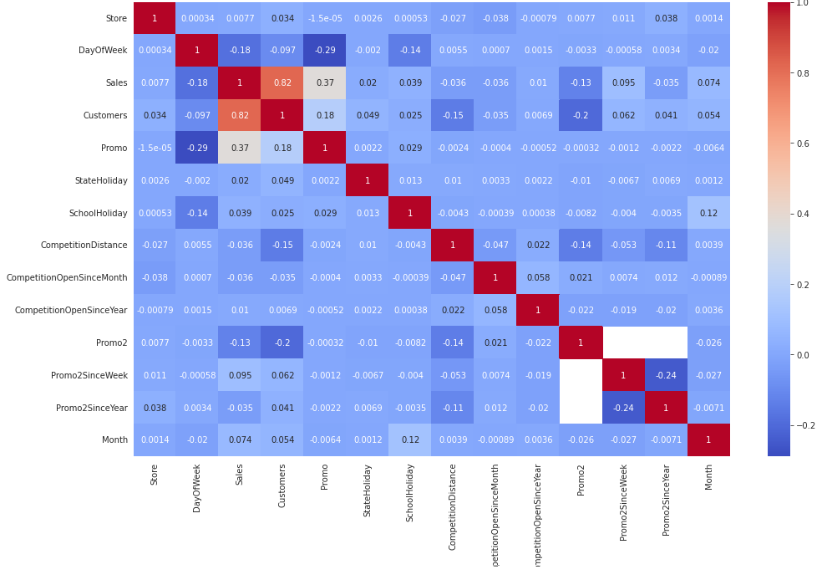


Here’s a distribution plot of the Sales column. The drop in sales indicates the 0 sales accounting for the stores temporarily closed due to refurbishment.

## Correlation:

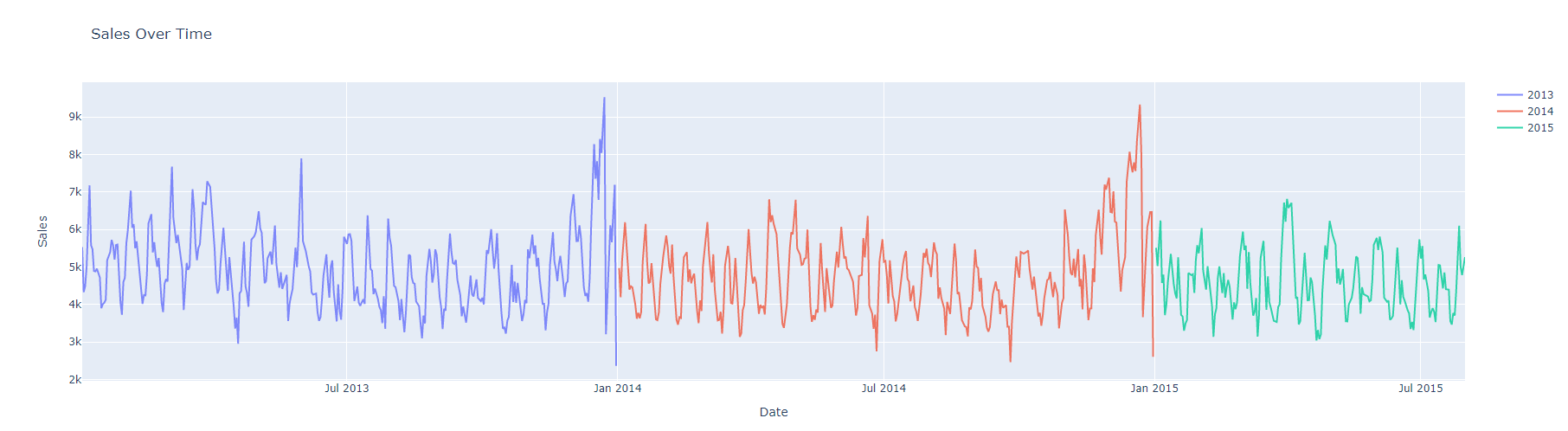
Correlation is a statistical term used to measure the degree to which two variables move in relation to each other. A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one variable move, either up or down, the other moves in the same direction. A perfect negative correlation means that two variables move in opposite directions, while a zero correlation implies 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 a positive correlation.
* State Holiday has a negative correlation suggesting that stores are mostly closed on state holidays indicating low sales.
* CompetitionDistance showing a 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.

**Sales Over Time**



Here’s a plot of Monthly Sales over the years. Sales rise by the end of the year before the holidays. Sales for 2014 went down there for a couple of months - July to September, indicating stores closed due to refurbishment.

## ****Sales Over Days Of Year:****

## 

## This graph shows sales over Days of a Year and the estimate of the number of customers the Stores has over the Year.

## This gives us the knowledge in which month the number of customers comes more frequently than usual and the Sales particular to those days

## Sales Over Weeks:

## 

## This graph shows sales over a week in the whole year and the estimate of the number of customers the Stores has over the Year.

## The graph gives us the knowledge of the weeks when the number of sales has been the most and when it is recorded as the lowest.

## Sales over Days of the week:

## 

## The graph shows the total sales over the week that the store had by combining the sales that happened on the days of the week over the whole year.

**Sales By Store Type:**

## 

## This shows the sales by store type that has been recorded and it can be seen that type b has recorded the most sales as compared to the other types of stores available.

## While all the other types of stores record the same type of sales.

**Sales vs. Number Of Customers**

## 

Sales and Customer scatter plot showed a direct positive relation between them with a few outliers.

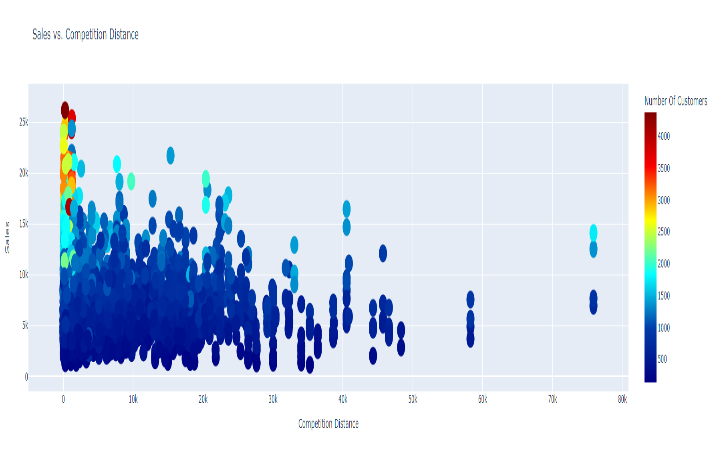
## Sales By Assortment:

## 

## Earlier, it was observed that **only store type b had all three kinds of assortment levels and the rest of the store types had two of them**. It seems that in some b type stores the products were different as compared to others because the revenue per store is significantly more than the others

The above bar plot shows that the store types a, c and d have only assortment levels a and c. On the other hand, store type b has all three kinds of assortment strategies, a reason why average sales were high for store type b stores.

**Sales vs. Competition Distance:**

****

From the above scatter plot it can be observed that mostly the competitor stores weren’t that far from each other and the stores densely located near each other saw more sales. This could indicate competition between busy locations vs remote locations.

## Sales Affected By School Holiday or Not?

## 

## As we can Conclude Sales are not at all affected by the school holidays. That means the number of customers coming is either adolescents or adults.

## Only 20% of the sales are affected but we can simply ignore the fact.

It is clear that most of the stores **remain closed**

**during State and School Holidays**

## 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 the biased nature of machine learning algorithms towards features with greater values and scale. The two techniques are:

Normalization: is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as

Min-Max scaling. [0,1]



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 the resultant distribution has a unit standard deviation. [-1,1]



**Both Standardization as well Normalization are used when we have more Numerical columns but since 70% of the columns are categorical columns and the Scaling technique does not affect these the   
best algorithm used in such scenarios is Decision tree and Random Forest. So, these techniques would not be used for this model.**

## Label encoding:

For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. We have categorical data integers encoded with us, but 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 and had to be Labeled encoded to continue.

# Modeling:

Factors affecting choosing the model:

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

## 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 case with time series data. Here, it’s important to use values at the rear of the dataset for testing and everything else for training.

The **30% of Data** was kept as a testing set and the rest of the historical data was used in the training set.

## Machine Learning Model –

Machine learning is the scientific study of algorithms and statistical models that computer systems use to

effectively perform a specific task without using explicit instructions, relying on patterns and inference

instead.

Building a model by learning the patterns of historical data with some relationship between data to make

a data-driven prediction.

**Types of Machine Learning**

• Supervised Learning

• Unsupervised Learning

• Reinforcement Learning

**Supervised learning**

In a supervised learning model, the algorithm learns on a labeled dataset, to generate reasonably

predictions for the response to new data.

(Forecasting outcome of new data)

• Regression

## • Classification

## Machine Learning models can be understood as a program that has been trained to find patterns within new data and make predictions. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response. First, these models are trained over a set of data, and then they are provided an algorithm to reason over data, extract the pattern from feed data and learn from those data. Once these models get trained, they can be used to predict the unseen dataset.

## Parameters that we need to check to evaluate the machine learning model that we are going to use. These parameters tell us about the model accuracy on training data, but it is also important to get a genuine and approximate result on unseen data otherwise Model is of no use.

## Parameters used to evaluate a model:

* Regression Model Score (Calculated using the scikit Learn Library)
* Out of Sample Test Score (Calculated using the scikit Learn Library)
* Training RMSE (Calculated using the scikit Learn Library)
* Testing RMSE (Calculated using the scikit Learn Library)
* Training MAPE (Calculated using the scikit Learn Library)
* Testing MAPE (Calculated using the scikit Learn Library)
* R2\_train (Calculated using the formula)
* R2\_test (Calculated using the formula)
* Adj\_r2\_train (Calculated using the formula)
* Adj\_r2\_test (Calculated using the formula

## Evaluation Metrics:

* Mean Absolute Positive Error (MAPE)- MAPE is a very simple metric that calculates the mean of the absolute difference between actual and predicted values.
* Root Mean Squared Error(RMSE)- It is a simple square root of mean squared error.
* Mean Squared Error(MSE)- Mean squared error states the mean of the squared difference between actual and predicted value.
* R Squared (R^2)- R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how well did your model perform. Hence, R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit. Its value ranges from 0 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 that while adding more data variance of data increases. Adjusted R^2 is adjusted for this disadvantage and shows the real value.

## Linear Regression:

Linear Regression tends to establish a relationship between a dependent variable(Y) and one or more

independent variable(X) by finding the best fit of the straight line.

## The equation for the Linear model is Y = mX+c, where m is the slope and c is the intercept

## The optimization algorithm used is Gradient Descent. We also performed different types of regularization techniques to prevent overfitting in the model.

## 

In the above diagram, the blue dots we see are the distribution of 'y' w.r.t 'x.' There is no straight line that runs through all the data points. So, the objective here is to fit the best fit of a straight line that will try to **minimize the error between the expected and actual value**.

## ****Linear Regression**** results were much better than other model with a sample test score Of 0.6648.

## ****Bayesian Ridge Regression:****

## **Overfitting happens when the model learns signal as well as noise in the training data and wouldn’t perform well on new/unseen data on which the model wasn't trained.**

## **To avoid overfitting, your model on training data like cross-validation sampling, reducing the number of features, purring, regularization, etc.**

## So to avoid overfitting, we perform Regularization

## 

## 

## Regularization adds the penalty as model complexity increases. The regularization parameter (lambda) penalizes all the parameters except intercept so that the model generalizes the data and won’t overfit.

## Ridge regression adds “squared magnitude of the coefficient" as a penalty term to the loss function. Here the box part in the above image represents the L2 regularization element/term

## ****Bayesian Ridge Regression**** results were much better than other model with a sample test score Of 0.6648.

## ****LARS Lasso Regression:****

The main objective of creating a model(training) is to make sure it fits the data properly and reduce the

loss. Sometimes the model that is trained which will fit the data but it may fail and give a poor

performance during analyzing of data (test data). This leads to overfitting. Regularization came to

overcome overfitting.

Lasso Regression (**Least Absolute Shrinkage and Selection Operator**) adds “Absolute value of **magnitude” of coefficient, as penalty term to the loss function**

## 

## Lasso Lars****Regression**** results were much better than other model with a sample test score Of 0.4687.

## 

## Decision Tree:

A decision tree is a type of supervised learning algorithm that can be used in classification as well as regressor problems. The input to a decision tree can be both continuous as well as categorical. The decision tree works on an if-then statement. A decision tree tries to solve a problem by using tree representation (Node and Leaf)

* Assumptions while creating a decision tree: Initially all the training set is considered as a root
* Feature values are preferred to be categorical, if continuous then they are discretized
* Records are distributed recursively on the basis of attribute values
* Which attributes are considered to be in root node or internal node is done by using a statistical approach.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches

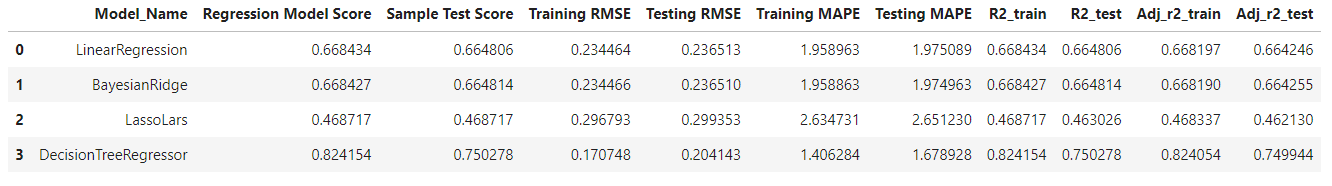


The results show that a simple decision tree is performing pretty well on the validation set but it has completely overfitted the train set with a sample test score of 0.75. It's better to have a much more generalized model for future data points.

Businesses prefer the model to be interpretable to understand the patterns and strategize accordingly unlike any scientific.

The facility where the results matter much more than interpretability.

If interpretability is important then sticking with tree-based algorithms when most of the features are categorical; is beneficial and using tuned Hyperparameters to grow the tree deep enough without overfitting.

****

## Random Forest:

Random forests are an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time. For regression tasks, the output of the random forest is the average of the results given by most trees.

In simple terms, a random forest builds multiple decision trees and merges them to get a more accurate and stable prediction.

Random Forest Regressor results were much better than other model with a sample test score

Of **0.831089**.

**Random Forest Hyperparameters:**

* max\_depth- The max\_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node
* min\_sample\_split- a parameter that tells the decision tree in a random forest the minimum required number of observations in any given node to split it.

The default value of the minimum\_sample\_split is assigned to 2. This means that if any terminal node has more than two observations and is not a pure node, we can split it further into subnodes.

* max\_leaf\_nodes- This hyperparameter sets a condition on the splitting of the nodes in the tree and hence restricts the growth of the tree. If after splitting we have more terminal nodes than the specified number of terminal nodes, it will stop the splitting and the tree will not grow further.
* min\_samples\_leaf- This Random Forest hyperparameter specifies the minimum number of samples that should be present in the leaf node after splitting a node.
* n\_estimators- the number of trees
* max\_sample (bootstrap sample)-The max\_samples hyperparameter determines what fraction of the original dataset is given to any individual tree.
* max\_features- This resembles the number of maximum features provided to each tree in a random forest.

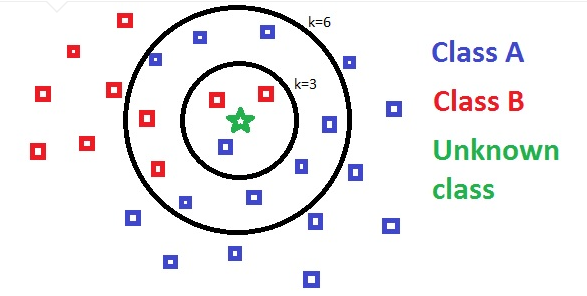
## ****K-Nearest Neighbors Regression:****

KNN regression is a lazy learning and non- parametric method that uses ‘feature similarity’ to predict the values of new data points. It calculates the distance between test data and each row of training data with the help of any of the methods namely: Euclidean, Manhattan, or Hamming distance. Assign the points which are nearest to it and approximate the result with the mean or mode values of its neighbors for regression or calcification respectively. The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation.

The method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.

.

## ****K-Nearest Neighbors Regression**** results were much better than other model with a sample test score Of 0.4772.



## Random Forest Hyperparameter Tuned Model :

Randomized search CV is used to perform a random search on hyperparameters. Randomized search CV uses a fit and score method, predict probe, decision\_func, transform, etc..,

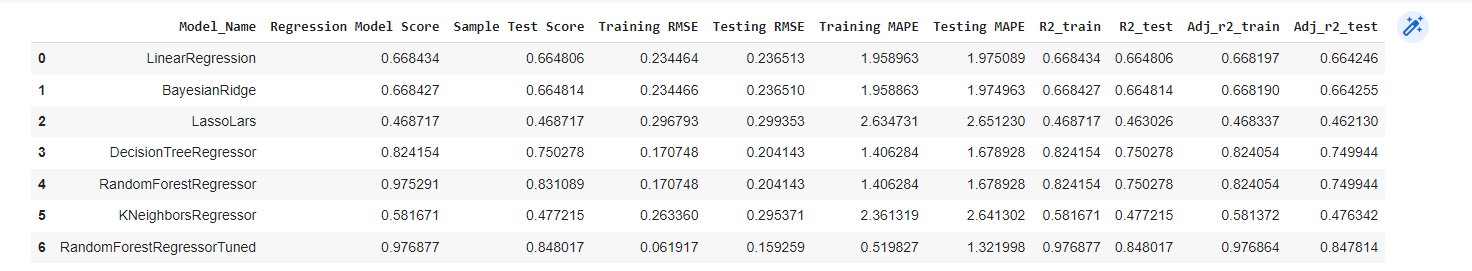
The parameters of the estimator used to apply these methods are optimized by a cross-validated search over parameter settings.

In contrast to GridSearchCV, not all parameter values are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions. The number of parameter settings that are tried is given by n\_iter.

## Random Forest Hyperparameter Tuned Model

The maximum Sample test Score was seen in the tuned Random Forest model with the value of **0.848**which was only **2.037 %** improved from a simple random forest model.

This indicates that all the trends and patterns that could be captured by these models without overfitting were done and the maximum level of performance achievable by the model was achieved.

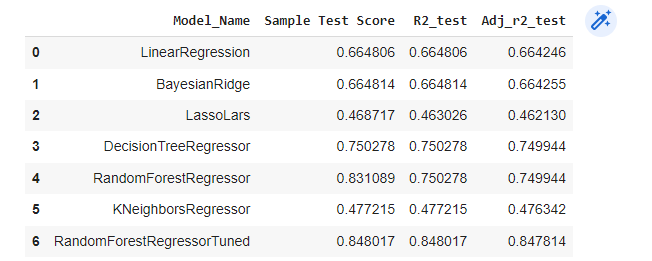


# Model Performance and Evaluation:

## 

## 

## 



Since the R^2 Score and the Adjusted\_r^2\_Score

are similar in two of the three cases and hence we

have to shift to the Sample Test Score.

## Random Forest vs Decision Tree

Model Performance

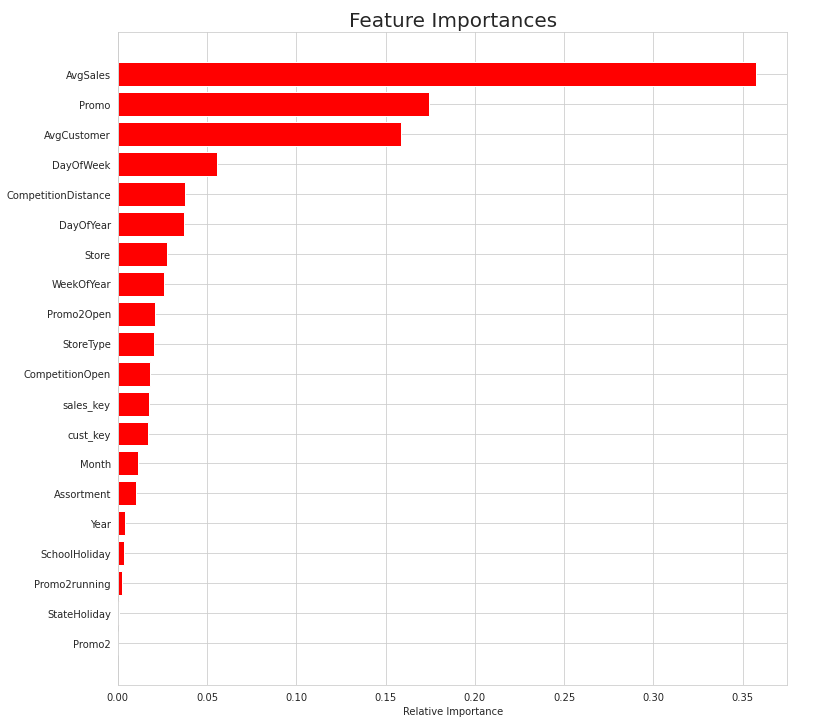
* Improvement of **10.771 %** was seen in **Random Forest against Decision Tree**

## Random Forest Tuned vs Random Forest Models

* The Sample Test Score was seen in tuned Random Forest model with the value **0.848** which was only **2.037%** improved from a simple random forest model.

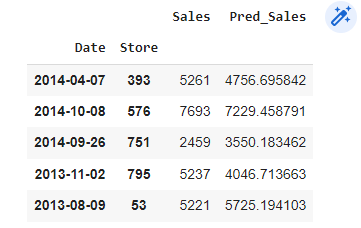
**Feature Importance:**

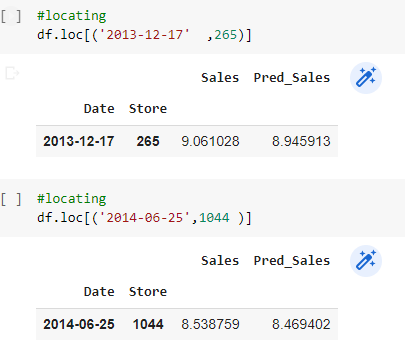
The most important features in predicting the Sales were Avg Sales, Promo, Avg Customers, Days of week and Competition Distance.



# Store-wise Sales Prediction:

Here is the actual sales values against the predictions which can be located date and store wise:





# Conclusion and Recommendations:

## 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 can explain **0.848593** **i.e. 84.85%** of the variations, and can predict the sales values in a good range.

Some important insights to draw from the analysis include:

* There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.
* The positive effect of promotion on Customers and Sales is observable.
* Most stores have competition distance within the range of 0 to 10 km and had more sales than stores far away, probably indicating competition in busy locations vs remote locations.
* Store type B though few had the highest sales average. The reasons include all three kinds of assortments especially assortment level b which is only available at type b stores and is open on Sundays as well.
* The outliers in the dataset showed justifiable behavior. The outliers were either of store type b or had a promotion going on that increased sales.
* Random Forest Tuned Model gave the best results and only 0.021% improvement was seen from the basic random forest model which indicates that all the trends and patterns that could be captured by these models without overfitting were done and the maximum level of performance achievable by the model was achieved.

**Recommendations:**

* More stores should be encouraged for promotion.
* Store type B should be increased in number.
* There's a seasonality involved, hence the stores should be encouraged to promote and take advantage of the holidays.

## Challenges:

* The major challenge would be the computational time and RAM needed to work upon such a dataset in a cloud environment.
* A huge amount of data needed to be dealt with while doing the project which is quite an important task and also even small inferences need to be kept in mind.
* As dataset was quite big enough which led to more computation time.

# References:

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