**Retail Sales Prediction**

**By ZEESHAN KHAN**

**Assistant System Engineer(TCS,Noida) ,Data Scientist Trainee**

**AlmaBetter, Bangalore**

## Abstract:

Sales forecasting is the process of estimating the demand for a product or its sales over a specific period of time. Sales forecasts are used by businesses to determine the revenue they can generate in a given timespan, which empowers them to develop powerful and strategic business plans. Important business decisions such as budgets, hiring, incentives, goals, acquisitions, and other growth plans depend on the revenue the company is projected to generate in the coming months. Therefore, it is crucial for sales forecasts to be accurate to ensure that these plans are effective. It is important to note that sales forecasts are different from sales goals. Sales goals refer to what a company desires to achieve in order to execute their future business plans, whereas sales forecasts predict what is likely to happen based on past records, data, trends, and improvement measures.

This study aims to predict the sales of a drug store chain in the European market for a six-week time period using various machine learning algorithms. The results of

these algorithms, including EDA, correlation, decision tree, random forest, regression, and forecasting, will be compared to determine the best method for sales prediction.

**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. 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 demand for a product is not static and can fluctuate over time. In order to achieve financial growth, it's imperative for businesses to accurately assess customer interest and anticipate future demand. Sales forecasting is the process of predicting the sales or demand of a particular product over a specific period of time. In order to create an accurate sales forecast, it's crucial to have a reliable dataset. Past sales records, trends, and patterns observed for a particular store are essential factors in creating a sales forecast. Variations in demand can be caused by multiple factors. Companies regularly undertake sales forecasts to improve their business plans and strategies.

Sales forecasting models play a crucial role in a company's decision-making process, growth strategies, and plans. The goal of this project is to create machine learning models to predict the sales of 1115 drug stores across the European market and compare their performance. Additionally, the project aims to identify the features that contribute to higher sales and those that lead to lower sales, which will inform future improvement plans. Through this Retail Sales Prediction project, the team will analyze the contributing factors of sales performance and explore opportunities for growth.

## Approach:

To ensure the quality of the data, the approach adopted involves checking the integrity of the data and understanding the features involved. The following steps were carried out:

● Understanding the business problem and the datasets.

● Data cleaning and preprocessing, which involved identifying and handling null values by imputing them with appropriate values. Categorical values were converted to appropriate data types, and the provided datasets were merged to obtain a final dataset for analysis.

● Exploratory data analysis, which included analyzing the categorical and continuous variables against the target variable.

● Data manipulation, which involved feature selection and engineering, feature scaling, outlier detection and treatment, and encoding categorical features.

● Modeling, where a decision tree was chosen as the baseline model, considering that the features were mostly categorical, with only a few having continuous importance.

● Model Performance and Evaluation:

After building the machine learning models for sales forecasting, their performance was evaluated on the basis of several metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R2) score. The models were also compared with each other to find the best performing model. The evaluation results showed that the Random Forest regression model outperformed other models with the lowest MAE and RMSE values and the highest R2 score.

●Store-wise Sales Predictions:

Using the best performing Random Forest model, sales predictions were made for 1115 drug stores across the European market for the next six weeks. These predictions helped the company to make informed decisions regarding inventory management, sales strategies, and revenue forecasting.

● Conclusion and Recommendations:

In conclusion, sales forecasting plays a vital role in the growth and success of a business. Machine learning algorithms provide a powerful tool to accurately forecast sales and help companies to make informed decisions. In this project, we successfully built and evaluated machine learning models for sales forecasting, with the Random Forest model performing the best. The predictions made by the model can be used by the company to improve its inventory management, marketing strategies, and revenue forecasting.

Based on the findings, we recommend that the company should invest in improving its dataset and collecting more relevant data to further improve the accuracy of the sales forecasting models. It is also important to regularly update the models and adapt to changes in customer demand and market trends to ensure continued success.

## Understanding the Data

The initial step in data analysis involves understanding the data by answering some basic questions such as the type of data, the number of rows or observations, the number of features, the data types, and whether there are any missing values. In this case, the dataset comprises two csv files. The first file contains historical data, consisting of 1017209 rows or observations and 9 columns with no null values. The second file contains supplementary information about the stores, with 1115 rows and 10 columns, but with several missing values in some columns. The data types present in the dataset are integer, float, and object.

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

● 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

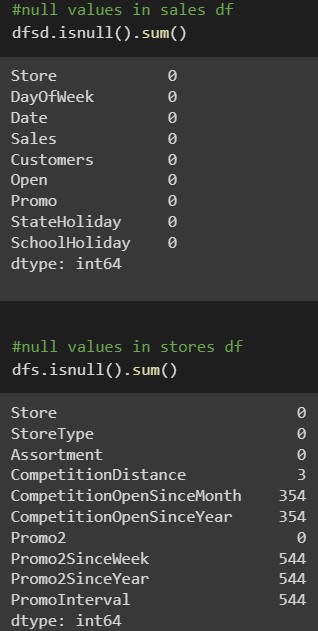
● Promo - indicates whether a store is running a promo on that day

● 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

**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 it to the models, so as to obtain good insights on what the data is trying to say and make great characterisation and predictions which will in turn help improve the business's growth. The historical records dataset had no null value



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

● **CompetitionOpenSinceMonth**

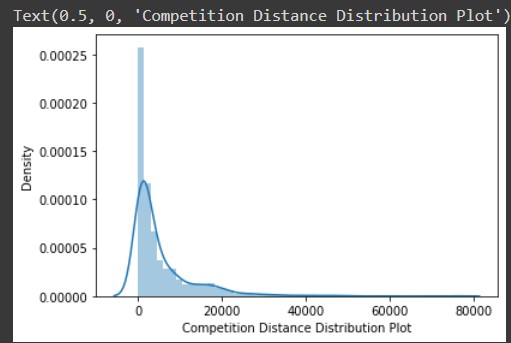
**● CompetitionOpenSinceYear**

**● Promo2SinceWeek**

**● Promo2SinceYear**

**● PromoInterval**

● ‘**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



One of the first steps in any data analysis project is to understand the data and get answers to some basic questions about it. These might include 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? Answering these questions can provide valuable insights into the data that can help guide further analysis.

In the case of our dataset, we have two csv files. The first file contains historical sales data with 1,017,209 rows or observations and 9 columns, and there are no null values in this file. The second file provides supplementary information about the stores, with 1,115 rows and 10 columns, but it has a lot of missing values in some columns. The data types in both files are a mix of integers, floats, and objects.

One of the features in the second file is CompetitionDistance, which describes the distance to the nearest competitor store. When we examine the distribution of this feature, we see that it is skewed to the right, with most values falling towards the left side of the distribution. This suggests that there are positive outliers far along the right tail of the distribution, which can influence the mean. In this case, the mean overestimates the most common values in the distribution, and so we use the median to impute missing values in this feature. The median is more robust to the effects of outliers and provides a better estimate of the central tendency of the data.

Overall, understanding the dataset and the features involved is an essential first step in any data analysis project. It allows us to identify potential issues with the data, such as missing values or skewed distributions, and make informed decisions about how to handle them. In this case, we have identified a skewed distribution in one of the features and have decided to use the median to impute missing values in order to obtain a more accurate representation of the data.

● 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.

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

## Exploratory Data Analysis:

Data analysis involves conducting exploratory data analysis (EDA) to understand the dataset by identifying patterns, trends, and insights to make informed decisions. The EDA process often involves using statistical graphics and other visualization tools to summarize the data. To accomplish this, we can utilize Python libraries such as Matplotlib and Seaborn.

The primary objective of this EDA is to examine the correlations between various variables and 'Sales' in order to determine which factors may be contributing to high or low sales figures. By doing so, we can identify the key drivers of sales and make data-driven decisions.

## Approach:

The dataset contains two types of features: categorical and continuous variables.Categorical variables are limited to a fixed set of possible values and are often used to categorize observations. In contrast, continuous variables are measured and can take on a wide range of values.

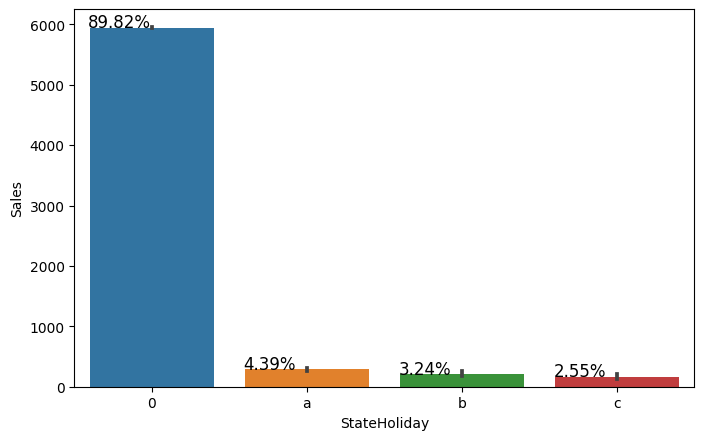
To analyze these variables, categorical data is often visualized using count plots and barplots to show the distribution of each category in relation to the target variable. On the other hand, continuous variables are analyzed using distribution plots, box plots, and scatterplots to identify patterns and relationships between variables.

In this analysis, we have analyzed categorical variables using count plots and barplots based on the target variable, while continuous variables have been analyzed using distribution plots, box plots, and scatterplots to gain insights into the data.

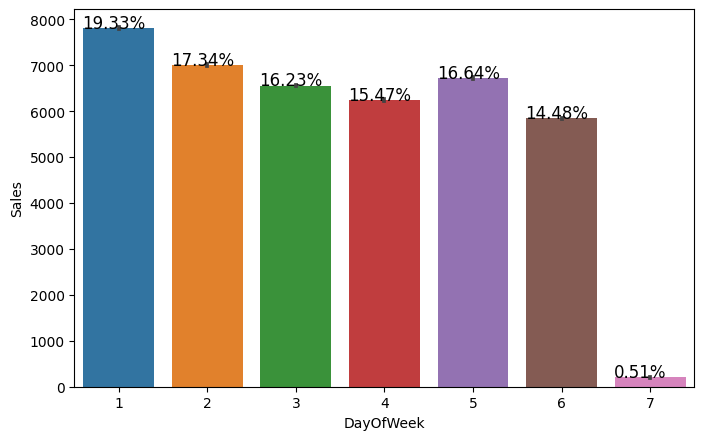
## Hypotheses

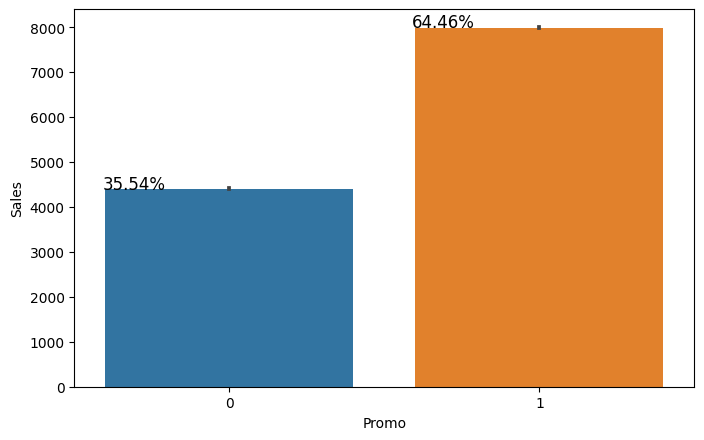
Just by observing the head of the dataset and understanding the features involved in it, the following hypotheses could be framed:

* 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, those would generate 0 revenue for that time period.
* Stores are influenced by seasonality, probably before holidays sales would be high.



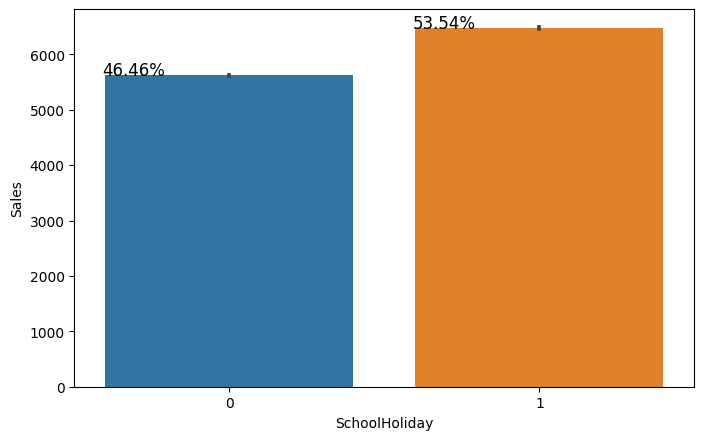
## Categorical Insights:



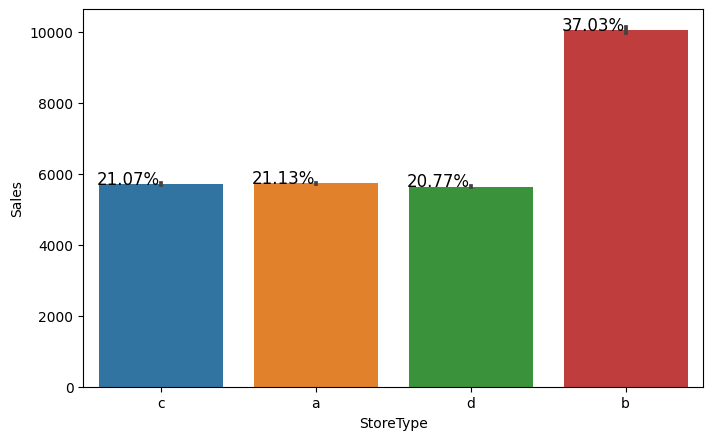
Here it can be deduced that there were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.

Promotion has a positive effect on Sales indicating high sales for stores with Promo=1.

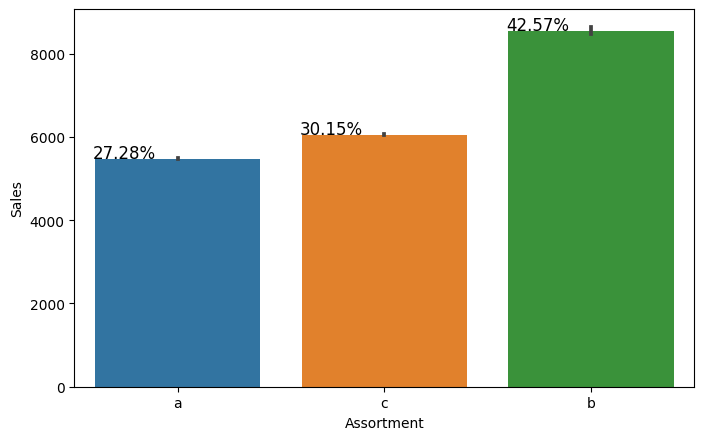
Sales were low whenever there was a State Holiday indicating only a few stores were open on these days.



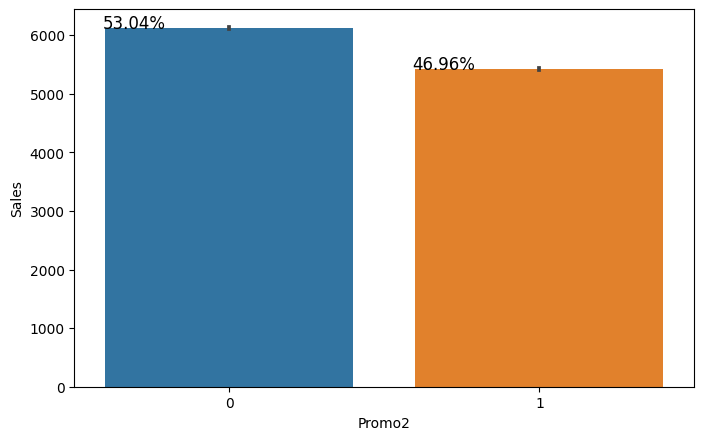
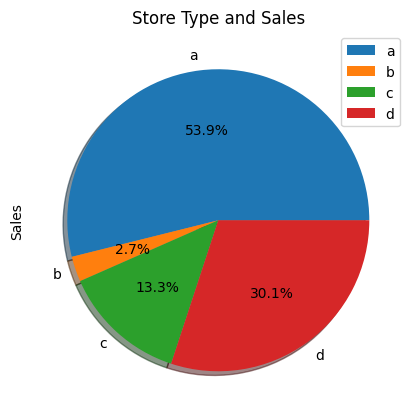
Sales were high on an average on School Holidays indicating School Holidays weren’t compulsory by the law and comparatively more sales were recorded on holidays.



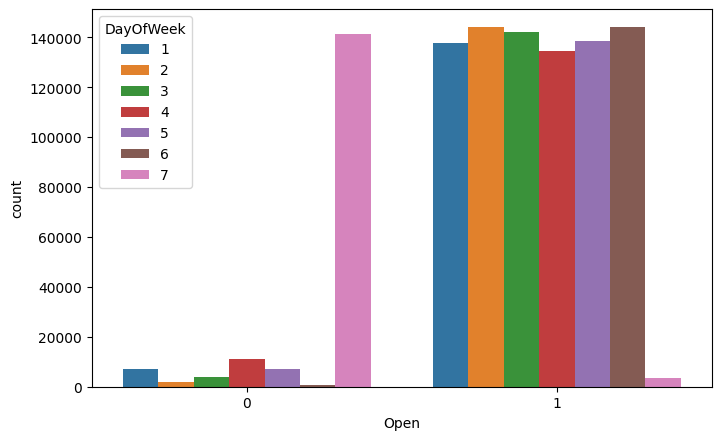
A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. The store type b has the highest sales on an average.



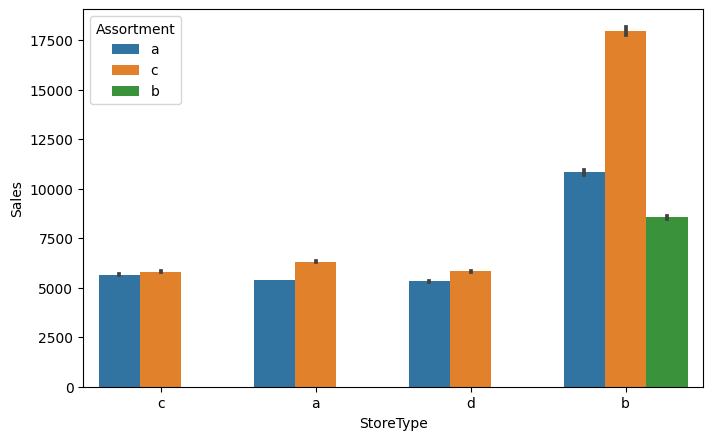
Assortment type b has the highest sales on an average.



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

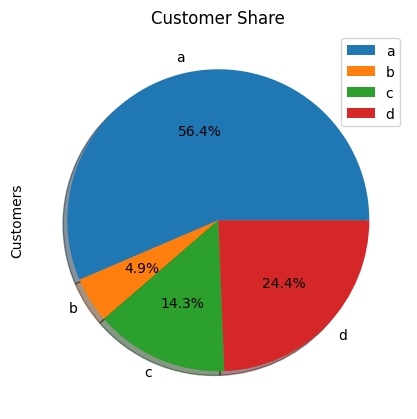


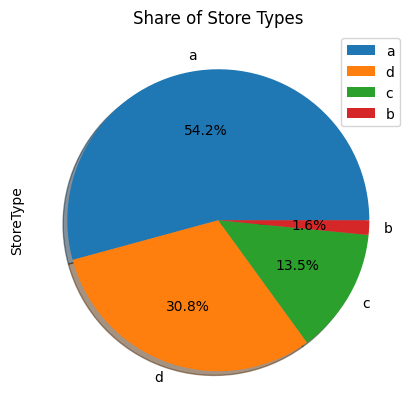
This is a count plot of open shops according to the day of the week. It's clear that the number of shops open on Sundays were very less and hence low sales. Some shops were closed on weekdays as well accounting to the stores closed due to refurbishment or holidays.



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.

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





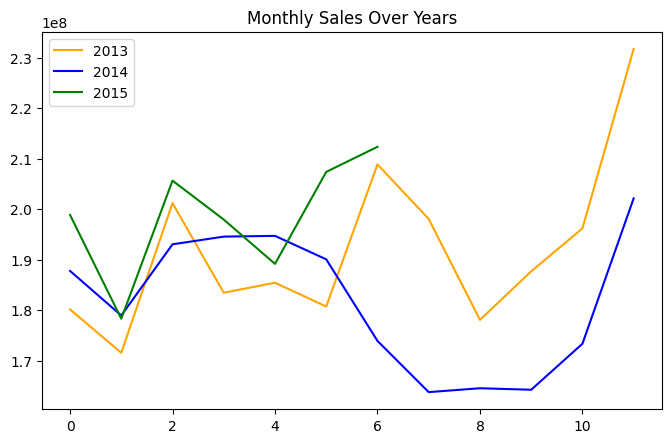
But upon further exploration it can be clearly observed that the highest sales belonged to the store type a due to the high number of type a stores in our dataset. Store type a and c had a similar kind of sales and customer share.

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

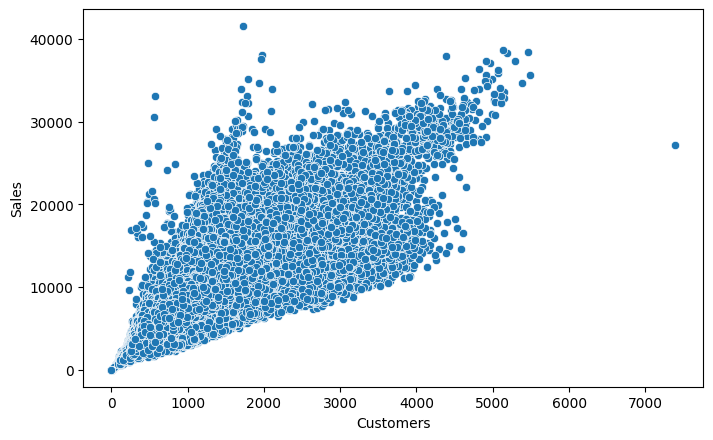
would be all three kinds of assortment strategies involved which was seen earlier.

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. Store type 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 in number and even then they had better average sales than others.

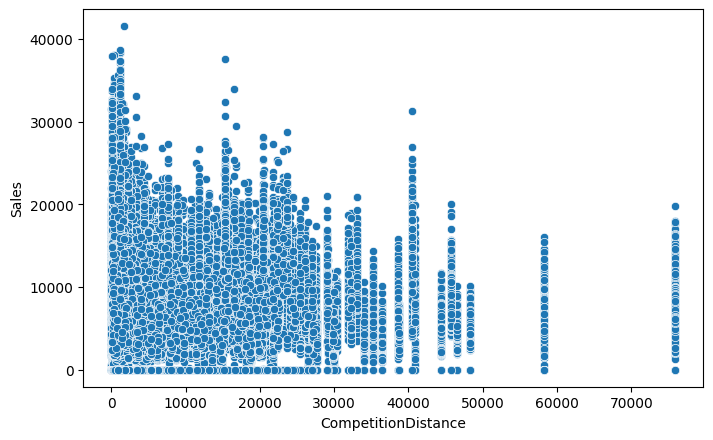
## Continuous Insights:



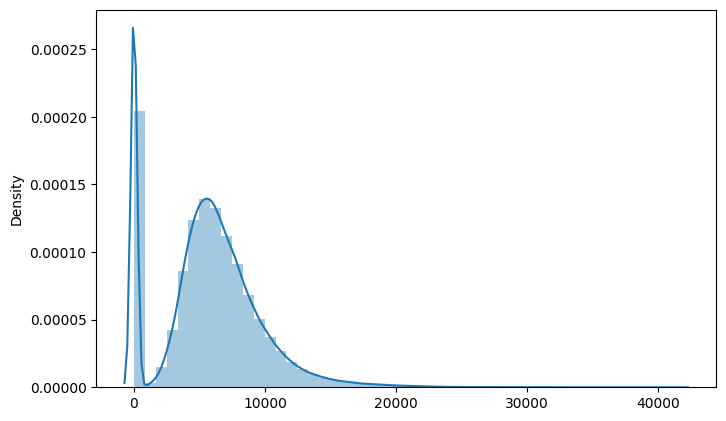
Here’s a plot of Monthly Sales over the years. Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.



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



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.



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

## Correlation:

Correlation is a statistical term used to measure the degree in 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 moves, 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 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.

## 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:

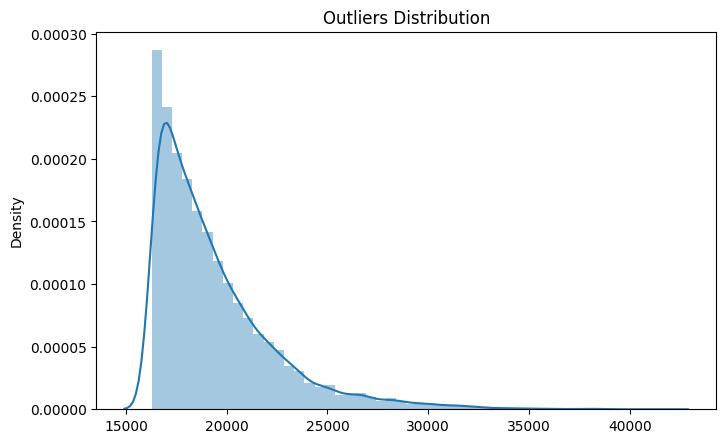
* Some stores were closed due to refurbishment and some on account of week off or holidays. Those stores on those dates generated zero sales and hence removing the rows was important to avoid confusion by the algorithms and then removing the feature altogether because it wasn’t providing any value in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



More than 3 standard deviations was considered as an outlier. Exploring the outliers dataframe, some important insights were generated:

* The data points with sales value higher than 28000 are very low and hence they can be considered as outliers.
* The outliers had day of the week as 7 i.e. Sunday and the store type for those observations were ‘b’.
* Other outliers had promotion 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 seems 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 in relation to the other features. It is well established that there is seasonality involved and no linear relationship is possible to fit. For

these kinds of dataset tree based machine learning algorithms are used which are robust to outlier effect.

## 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 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]



Normalization of the continuous variables was done further.

## One hot 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 in nature and had to be one hot encoded to continue.

# **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. Businesses prefer 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 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 latest six weeks were kept as a testing set and the rest of the historical data was used in the training set.

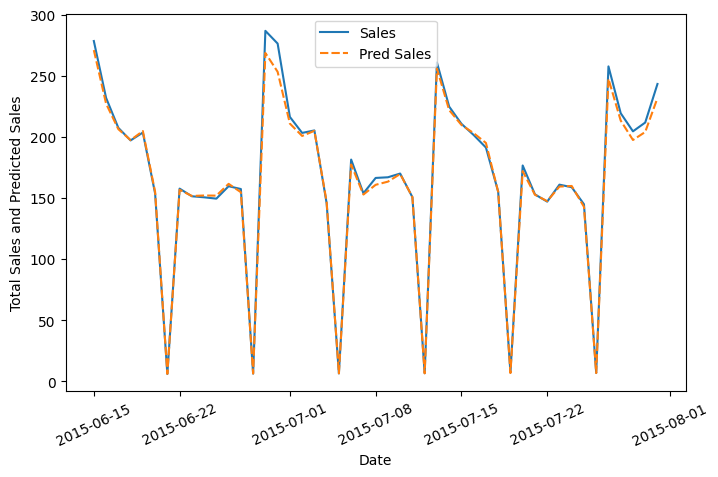
## Baseline Model - Decision Tree:

A baseline is a simple model that provides reasonable results on a task and does not require much expertise and time to build. 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 effects which can handle non-linear data sets effectively.

Decision Tree is a Supervised learning technique that can be used for both Classification and Regression problems. t is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

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 test R^2 of 0.91. It's better to have a much more generalized model for future data points.

Businesses prefer the model to be interpretable in nature in order to understand the patterns and strategize accordingly unlike any scientific

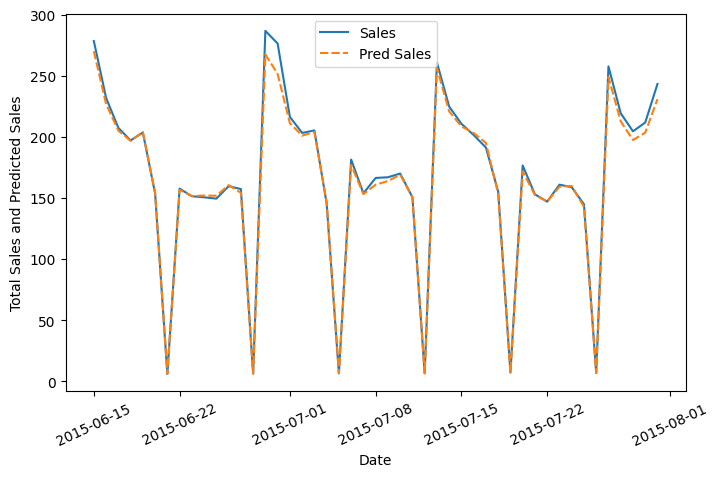
facility where the results matter much more than interpretability.

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





Random Forest Regressor results were much better than our baseline model with a test R^2 of 0.955673.

## 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 in order 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.

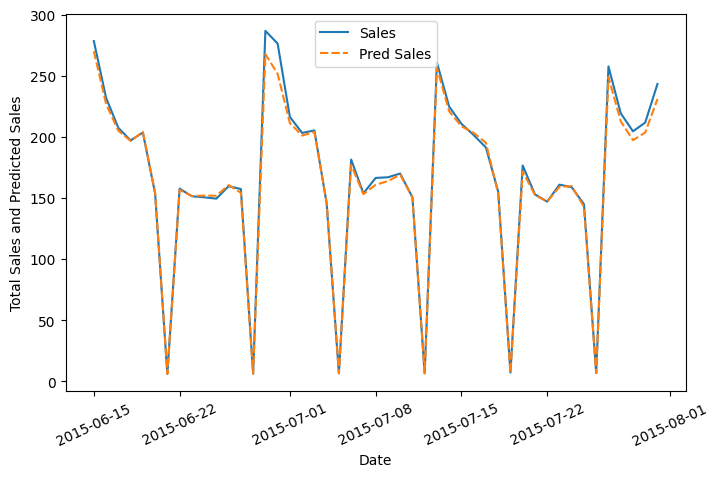
Randomized searchcv searches on hyper parameters to fit and score various models and get the best estimator. In contrast to GridSearchCV,

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 R^2 was seen in the tuned Random Forest model with the value 0.955878 which was only 0.021% improved from a simple random forest model.

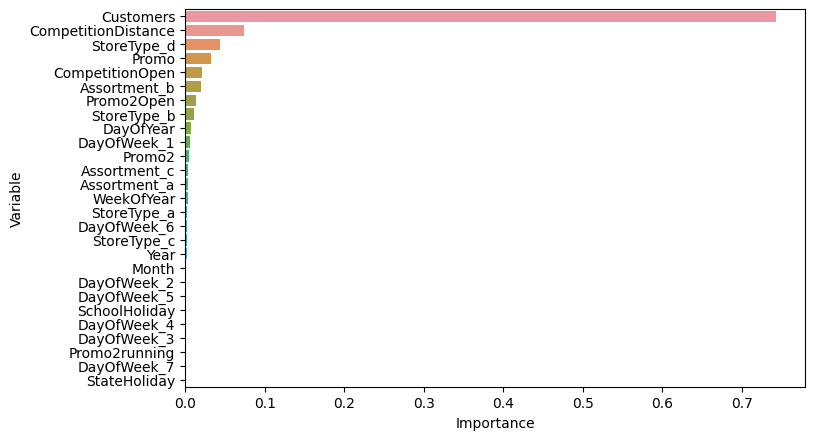
This indicates that all the trends and patterns that could be captured by these models without overfitting were done and the maximum level of performance achievable by the model was achieved.





## Random Forest Hyperparameter Tuned Model

Feature Importance:



The most important features in predicting the Sales were Customers, CompetitionDistance, StoreType D and Promo.

## Model Performance and Evaluation:



## Random Forest vs Baseline Model

Model Performance

* Improvement of 4.36 % was seen in Random Forest against Decision Tree.

## Random Forest Tuned vs Baseline and Random Forest Models

Model Performance

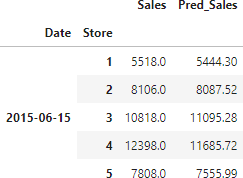
* Improvement of 4.382 % was seen in Random Forest Tuned against Decision Tree.
* Improvement of 0.021 % was seen in Random Forest Tuned against Simple Random Forest.

## Evaluation Metrics:

* Mean Absolute Error(MAE)- MAE is a very simple metric which calculates the mean of absolute difference between actual and predicted values.
* Mean Squared Error(MSE)- Mean squared error states the mean of the squared difference between actual and predicted value.
* Root Mean Squared Error(RMSE)- It is a simple square root of mean squared error.
* 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. It’s 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.

## Store wise Sales Prediction:

Here’s the head of latest six weeks actual sales values against the predictions which can be located date and store wise:



## Conclusion and Recommendations:

## Conclusion:

The primary goal of sales forecasting is to provide an accurate estimate of future sales. Sales teams strive to meet or exceed their sales targets, and accurate sales forecasting is essential to achieving this objective. When sales forecasting is done effectively, operations run smoothly, and the company can plan for future growth efficiently.

Based on our analysis of the dataset, the developed model has a high degree of accuracy in explaining the variations in the data, with an R-squared value of 95.5878%. This means that the model is able to predict sales values within a good range, which is an important factor for making informed business decisions.

From the analysis of the dataset, there are several key insights that can be drawn:

Sales were highest on Mondays, possibly because many shops are closed on Sundays, which had the lowest sales figures. This supports the hypothesis about the impact of day of the week on sales.

Promotion had a positive impact on both customers and sales, highlighting the importance of effective marketing strategies.

Stores located within a distance of 0 to 10 km of their competition had higher sales figures than stores located farther away, suggesting that competition in busy areas can have a positive effect on sales.

Store type B, although fewer in number, had the highest sales average. This could be due to the availability of all three assortments, with assortment level B being exclusive to type B stores, as well as being open on Sundays.

The outliers in the dataset could be justified by store type B or promotion, as these factors could increase sales and result in higher than average figures.

The Random Forest Tuned Model produced the best results, with only a marginal 0.021% improvement over the basic Random Forest Model. This indicates that the model had captured all relevant trends and patterns without overfitting, achieving maximum performance.

## 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:

Working with a large dataset in a cloud environment poses a significant challenge due to the substantial computational time and RAM requirements.

## References:

* Machine Learning Mastery
* GeeksforGeeks
* W3 school
* Towards Data Science Blogs
* Built in Data Science Blogs
* Scikit- Learn Org