Retail Sales Prediction

# By Annayan Bose, Prabhat Patel

# Data Science Trainee

# AlmaBetter, Bangalore

**Abstract:**

Sales forecasting is the technique of predicting demand or sales of a specific product over a predetermined time frame.

Businesses use sales forecasts to estimate the amount of income they will bring in over a specific period of time so they may create strong and effective business plans. The income the firm expects to generate in the upcoming months has an impact on crucial decisions like budgets, hiring, incentives, objectives, acquisitions, and numerous other growth plans, thus it's critical that these projections be accurate for the plans to be as successful as they are intended to be.

The sales goals that a corporation has differ from the sales estimates. A corporation needs to achieve its sales targets in order to carry out its long-term business strategies. Sales predictions, on the other hand, predict what will happen based on historical data, trends, and other improvement initiatives implemented.

In this study, machine learning algorithms are used to anticipate sales for a chain of pharmacies in the European market during a six-week period.

# 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:

A product's interest continues to shift from time to time. No company can focus on its financial expansion without carefully gauging customer interest and product demand in the future.

Sales forecasting is the technique of predicting demand or sales of a specific product over a predetermined time frame.

It is crucial to obtain a quality dataset in order to produce a reliable sales prediction. Forecasts heavily rely on historical data, trends, and patterns associated with sales at a specific store. Numerous factors could be to blame for the variances. Talking from a business’s point of view, these sales forecast

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

# Approach:

Here, the strategy is to first examine the data's integrity before comprehending the relevant properties. 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 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 categorical features.
* **Modeling**- The baseline model- Decision tree was chosen considering our features were mostly categorical with few having continuous importance.

## Model Performance and Evaluation

* **Store wise Sales Predictions**

# Conclusion and Recommendations

# Understanding the Data:

Understanding the data and obtaining the answers to some fundamental questions, such as What is the data about, is the first stage. What is the number of rows or observations in it? How many features does it have? What types of data are there? Exist any missing values here? Aside from anything else that might be pertinent to our study. Before moving on, let's just take a moment to comprehend the dataset and the associated terminologies.

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 of 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
* **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.

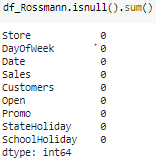
**Data Cleaning and Preprocessing:**

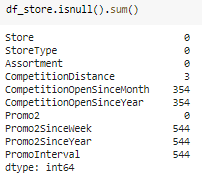
The ability to handle missing values is crucial for the data analysis process. We might decide to remove rows with missing values if there are very few of them in relation to the size of

missing values. Otherwise, it is better to replace them with appropriate values.

Before feeding the data to the models, it is important to check and handle these values in order to gain a clear understanding of what the data is trying to tell us and to develop excellent characterizations and forecasts that will boost the growth of the company.

The historical records dataset had no null values.

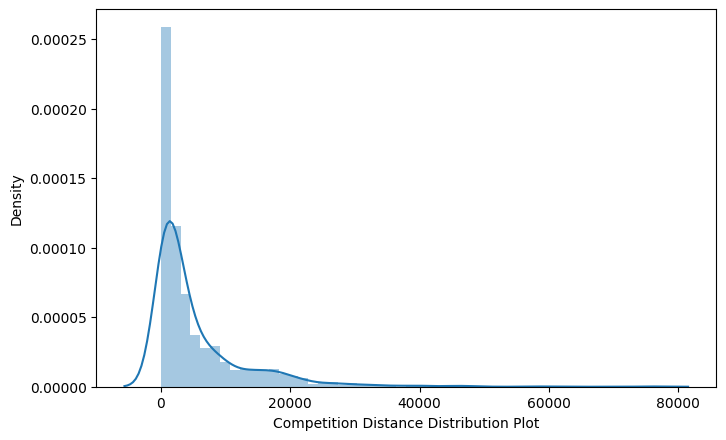




The above dataset had a lot of nulls in the following columns:

* 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.



It seems like most of the values of the CompetitionDistance 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 as positive skewed distribution which essentially suggests that there are positive outliers far along which influences the mean.

The majority of the CompetitionDistance values in the column appear to be in the range of 0 to 10 kilometres. Therefore, with an asymmetrical distribution, the longer tail drags the mean away from the most frequent values. Greater than the median is the mean. Because it is more resistant to the effects of outliers and underestimates the most frequent values in the distribution, the median is chosen in this situation to impute the missing values in this characteristic.

* 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.

# 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

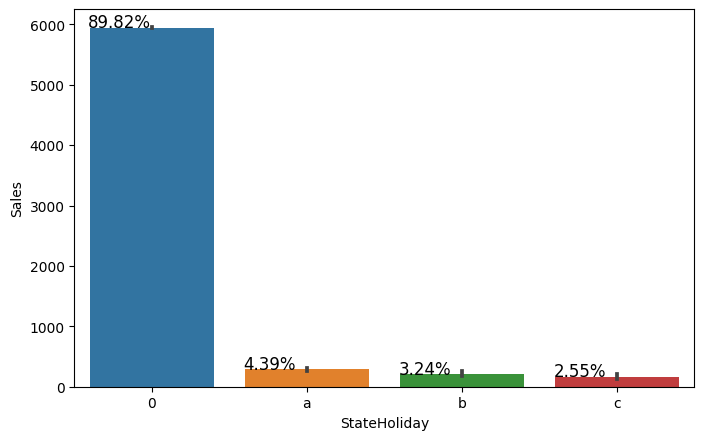
values putting a particular category to the observation.

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

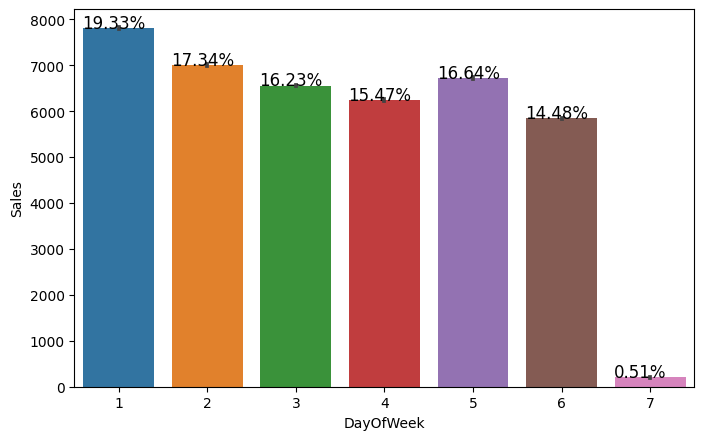
Both of them are analyzed separately. Categorical data is usually analyzed through count plots and barplots in accordance with the target variable and that is what is done here too. On the other hand Numeric or Continuous variables were analyzed through distribution plots, box plots and scatterplots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

## Preliminary Analysis:

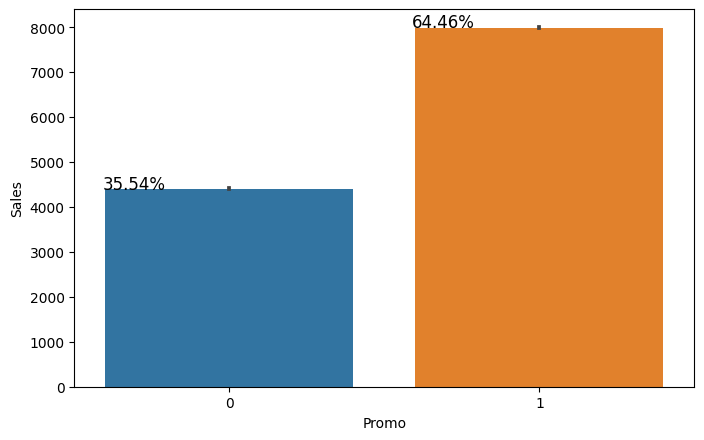
Just by observing the head of the dataset and understanding the features involved in it, the following preliminary assumptions could be framed:

* 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:

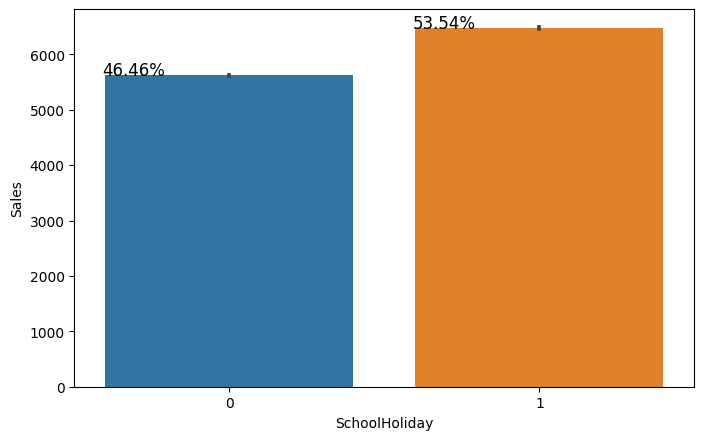


Here, it can be inferred that Monday saw increased sales, perhaps as a result of the fact that stores are often closed on Sundays, which saw the lowest sales during the week. This supports the theory on this trait.

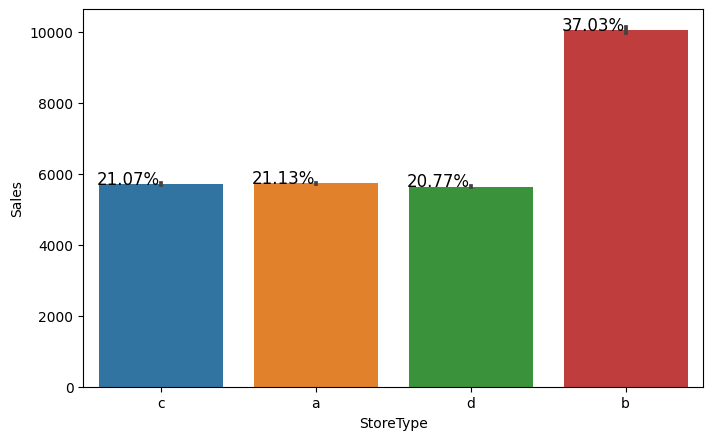


Promotion has a positive effect on Sales indicating high sales for stores with Promo=1.

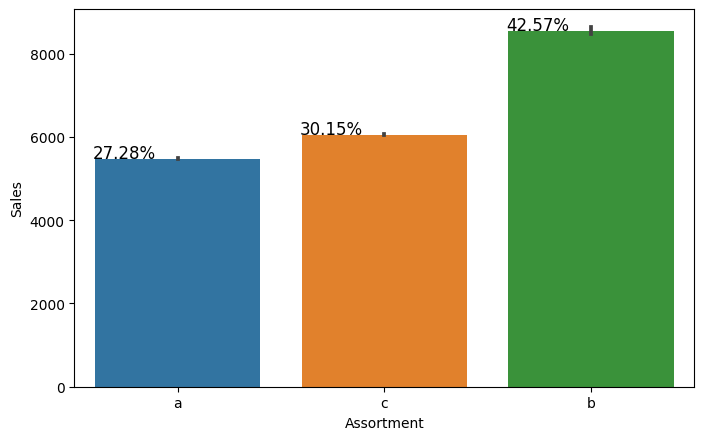
When there was a State Holiday, sales were weak, indicating that few stores were open.



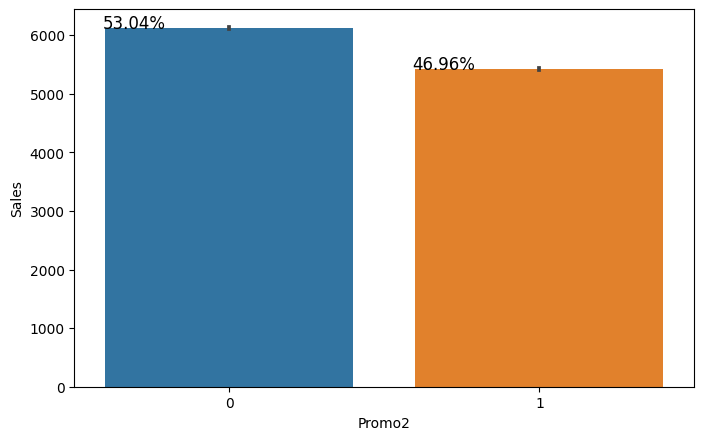
Sales were generally high during school breaks, showing that these breaks were not required by law and that proportionately more sales were made during these times.



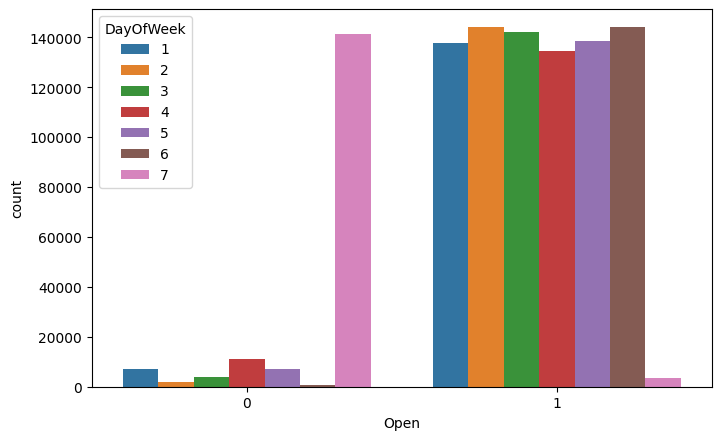
A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. The store type b has the highest average sales on an average.



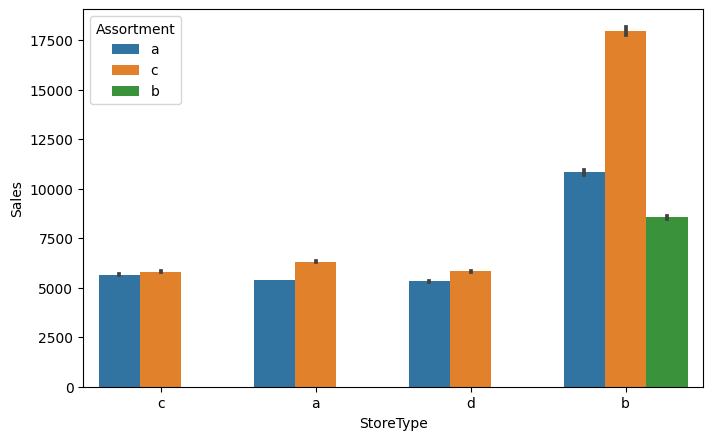
Assortment type b has the highest sales on an average.



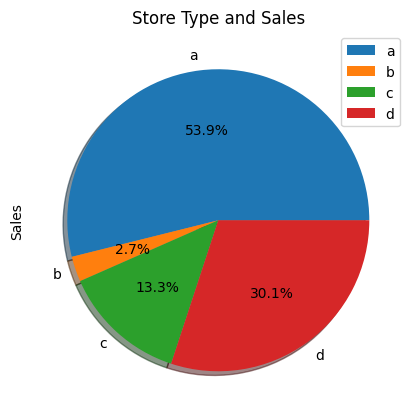
Without Promo2, sales were somewhat higher, indicating that many stores are not taking part in the promotion.



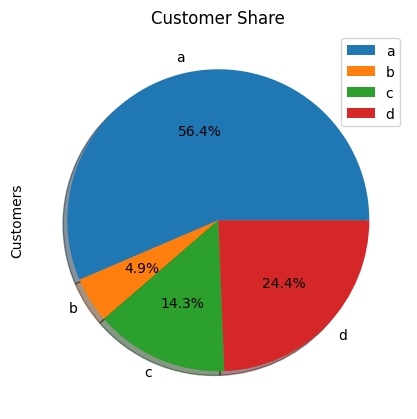
This is a count plot of open stores broken down by weekday. There were obviously very few stores open on Sundays, which resulted in minimal sales. Due to the stores being closed for renovation, some businesses were also closed on weekdays.

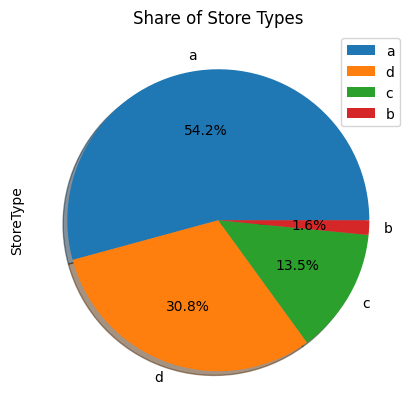


The accompanying bar plot demonstrates that only levels a and c of assortment are available at store types a, c, and d. However, store type b stores have all three types of assortment techniques, which is why their average sales were high.



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.





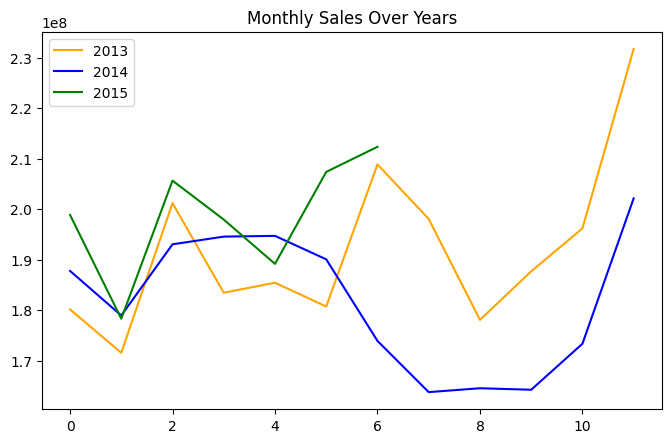
But after more investigation, it became evident that the store type a had the largest sales since there were so many type a retailers in our dataset. Sales and customer share for stores of types A and C were comparable.

A noteworthy finding is that the store type b with the greatest average sales and per-store income generation appears to be in good health

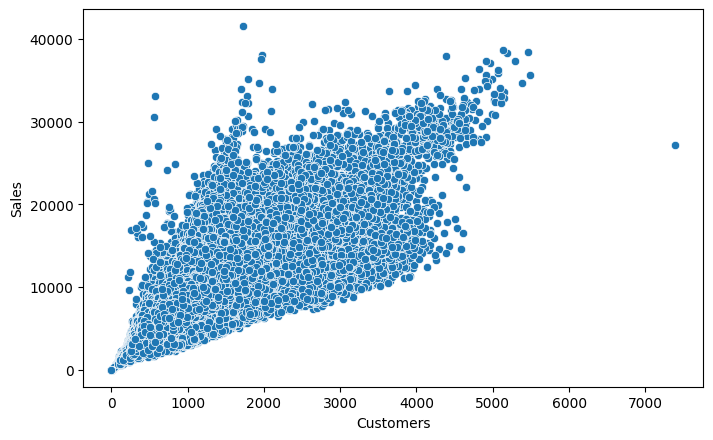
and reason for that would be all three kinds of assortment strategies involved which was seen earlier.

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.

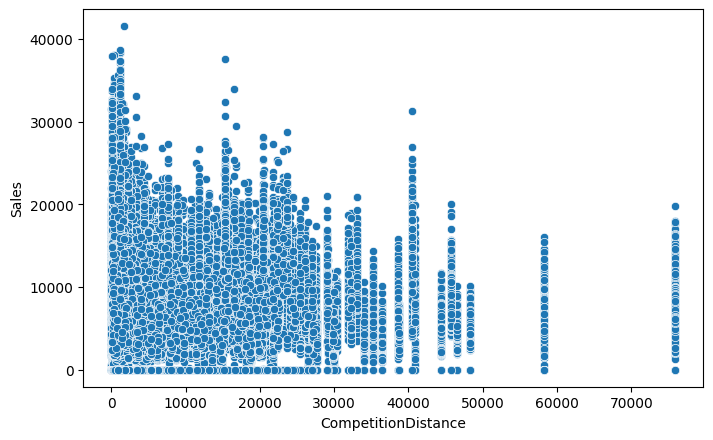
## Continuous Insights:



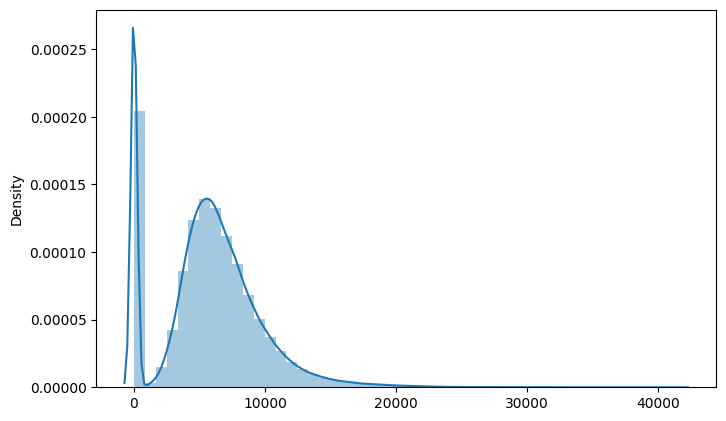
Here’s a plot of Monthly Sales over the years. Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.



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.

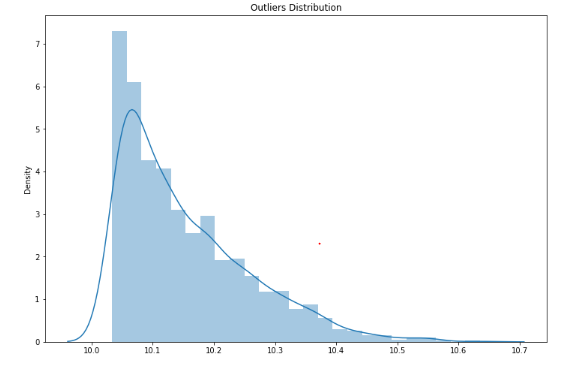
.

## 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:

* We observed that lot of stores are participating in promo and hence having high sales. So these are valid outliers..
* Also a lot of stores these are opened during Sundays as well which is possible reason for high sales.
* Some of the stores are neither running any promo and nor they have their stores opened on Sunday but still having high sales because they have very large number of customers visiting the stores.
* 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]



**Standardization** 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:

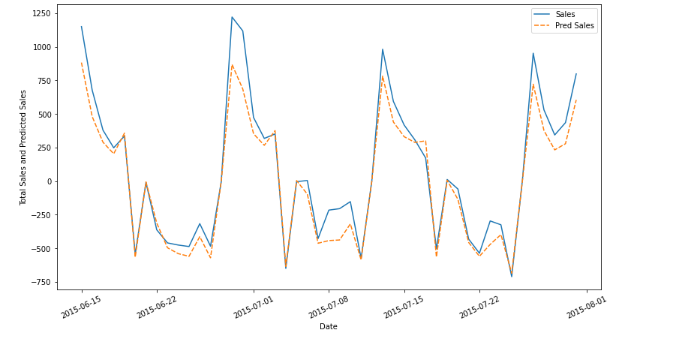
In machine learning, train/test split splits the data randomly, as there’s no dependence from one observation to the other. That’s not the case with time series data. Here, it’s important to use values at the rear of the dataset for testing and everything else for training.

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

**Linear Regression:**

The Linear regression model score is 73.83% on test dataset. As our dataset have most of the categorical columns with some continuous features such as customers and competition distance due to which linear regression is not quite good in terms of predicting the output.

* Overall if we see the graph of predicted sales vs actual sales most of the points predicted are somewhat close to actual values of sales but there are few points where it is not able to predict well. This problem may be resolved by applying Decision Tree model.





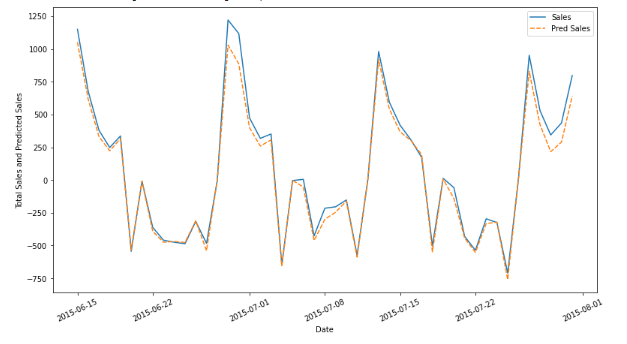
## Baseline Model - Decision Tree:

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.

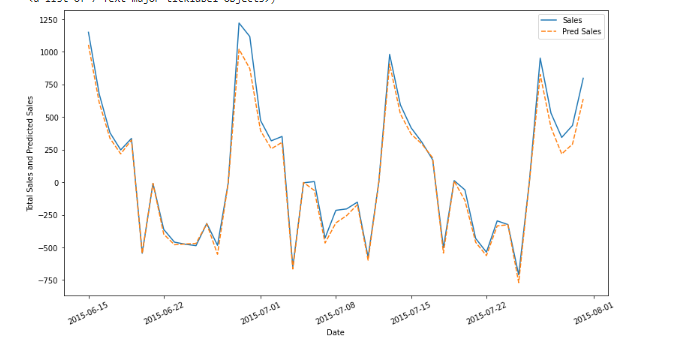




Decision tree is 92.97 percent efficient in predicting the sales. When we have mostly categorical columns in our dataset then decision tree is good choice and from our predicted vs actual sales graph also we can say that Decision tree has predicted most of the points very close to actual sales. Although the test score is 92.97% but if we see the train score is 99.6% so we'll try to reduce increase the test score by applying the Random forest regressor. The baseline model- Decision tree was chosen considering our features were mostly categorical with few having continuous importance. The above results show that a simple decision tree is performing pretty well on the validation set but it has completely overfitted the train set. It's better to have a much more generalized model for future data points.

## 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.



Random forest regressor is 96.23% effecient in predicting the output.

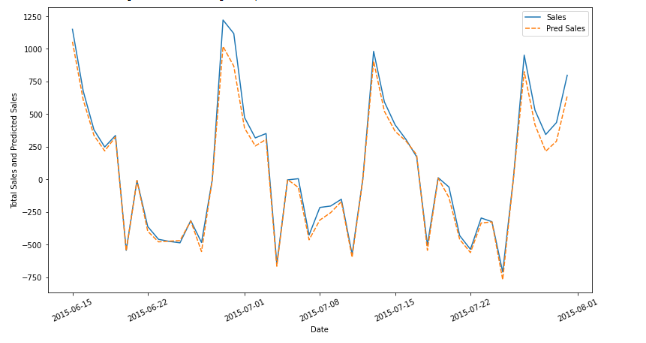
As from the graph of predicted vs actual sales also we can see most of the datapoints the predicted value is follwing the actual sales very closely.

.

## Random Forest Hyperparameter Tuned Model :

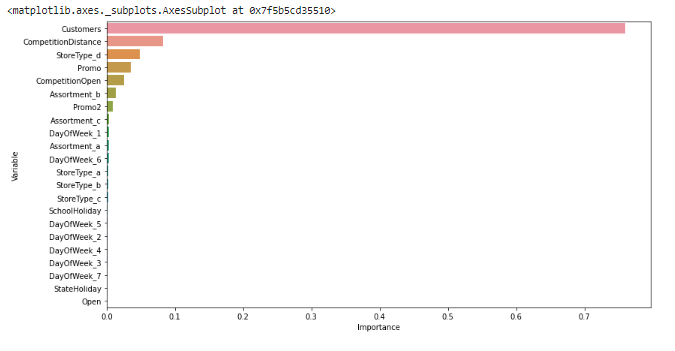
The maximum R^2 was seen in the tuned Random Forest model with the value 0.9624 which was only 0.023% improved from a simple random forest model.

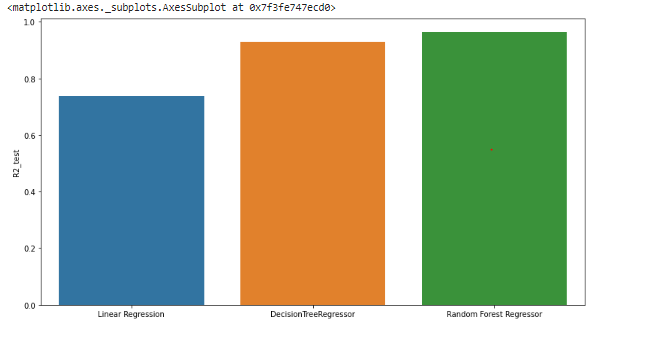
This indicates that all the trends and patterns that could be captured by these models without overfitting were done and the maximum level of performance achievable by the model was achieved.



## 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 3.508 % was seen in Random Forest against Decision Tree.

## Random Forest Tuned vs Baseline and Random Forest Models

**Model Performance**

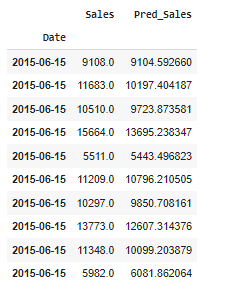
* Improvement of 3.532 % was seen in Random Forest Tuned against Decision Tree.
* Improvement of 0.023 % was seen in Random Forest Tuned against Simple Random Forest.

## 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:

****

**As we can see from above rows the prediction are very close to actual values.**

# Conclusion and Recommendations:

## Conclusion:

Some important conclusions drawn from the analysis are as follows:

* There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week.
* The positive effect of promotion on Customers and Sales is observable.
* Most stores have competition distance within the range of 0 to 10 kms and had more sales than stores far away probably indicating competition in busy locations vs remote locations.
* 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.
* The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.
* Decision tree was chosen as baseline model considering our features were mostly categorical with few having continuous importance.
* Random Forest shows improvement of 3.508% as compared to Decision tree.
* Random Forest Tuned Model gave the best results and only 0.023% 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 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.

# References:

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