Retail Sales Prediction

By
Neel Naik
AlmaBetter, Bangalore

Abstract:

Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time. 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 what a company wants to happen to execute their future plans for the business. On the other hand sales forecasts are what is going to happen on the basis of past records, data, trends and various improvement measures taken. The work here predicts the sales for a drug store chain in the European market for a time period of 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. 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 interest for a product continues to change occasionally. No business can work on its monetary growth without assessing client interest and future demand of items precisely. Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time. For a good sales forecast, it is extremely important to get a good dataset as well. Forecasts heavily depend on the past records, trends and patterns observed for sales of a particular store. The variations could be due to a number of 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 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:

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 categorical features.
- Modeling- The baseline modelDecision 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:

First step involved is 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 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)

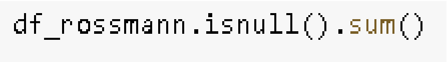
- **Customers -** the number of customers on a given day
- **Open -** an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday -** indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays 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
- **Promo2 Since[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.

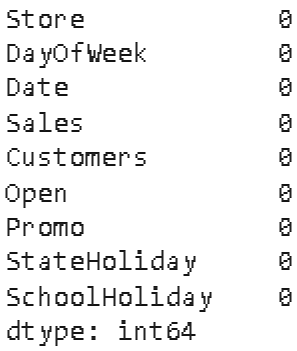
Data Cleaning and Preprocessing:

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.

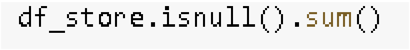
The historical records dataset had no null values.



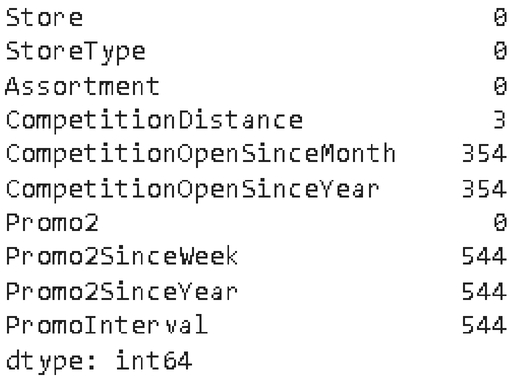
```
df_rossmann.isnull().sum()
```



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



```
df_store.isnull().sum()
```

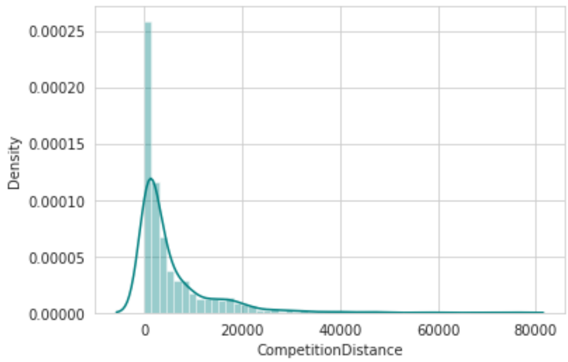


```
Store                          0
StoreType                      0
Assortment                     0
CompetitionDistance            3
CompetitionOpenSinceMonth    354
CompetitionOpenSinceYear     354
Promo2                         0
Promo2SinceWeek              544
Promo2SinceYear              544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

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

- 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. It seems like most of the values of the CompetitionDistance in the column are between 0-10kms. Consequently, the longer 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 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.

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.

Hypotheses:

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

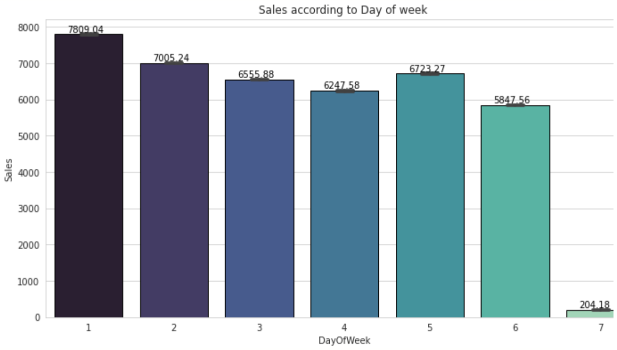
* Due to the high number of public holidays in December, sales will be at their highest.

* Due to weekends, sales ought to be at their peak on Saturday or Sunday.

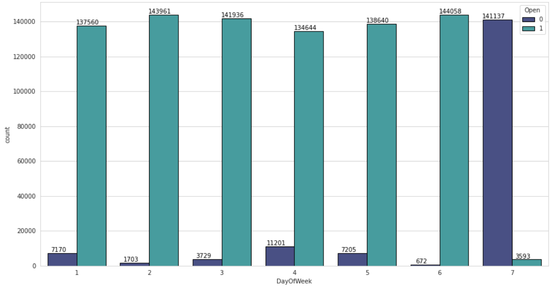
* Sales and promotion ought to be closely related in a favorable way.

- Due to its small number of stores, Store B will have the lowest sales.
- The aggregate sales are increased when competitors are close to one another.

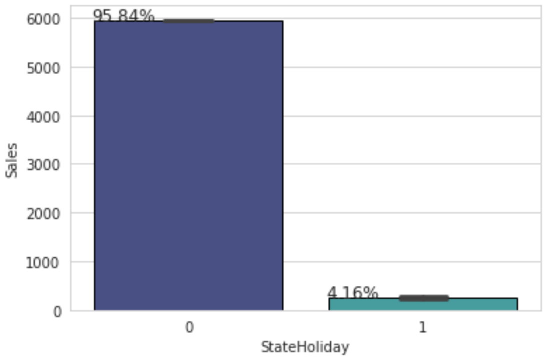
Categorical Insights:



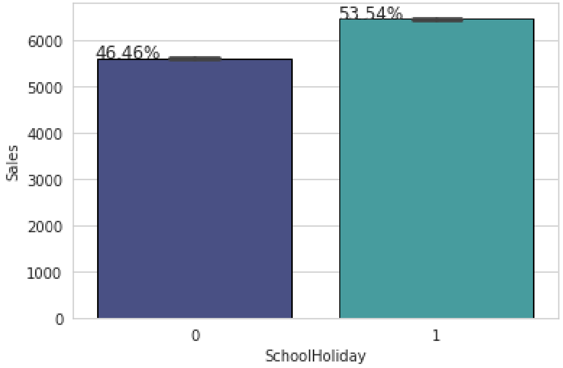
Here it can be deduced that there were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.



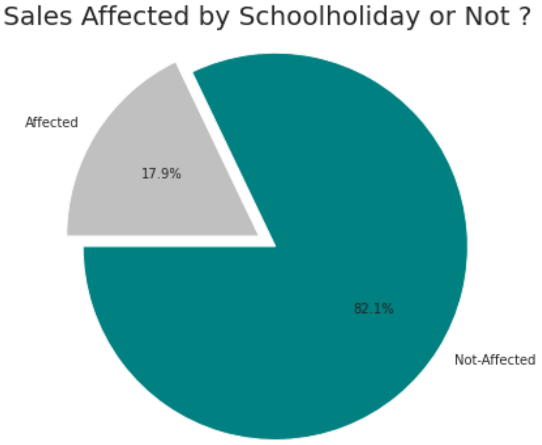
Based on the day of the week, the total number of store opening and closing days.



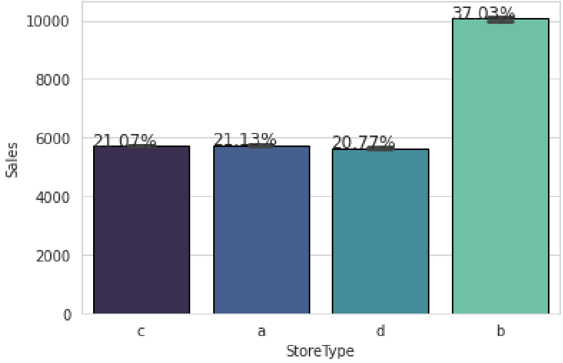
On state holidays most stores are closed resulting in 0 sales.



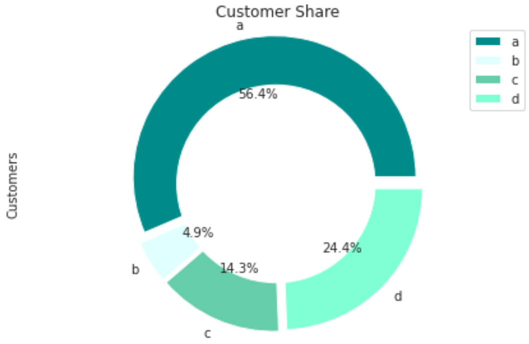
Most but not all school holidays are public holidays resulting in approximately half days with 0 sales and other half with some sales.



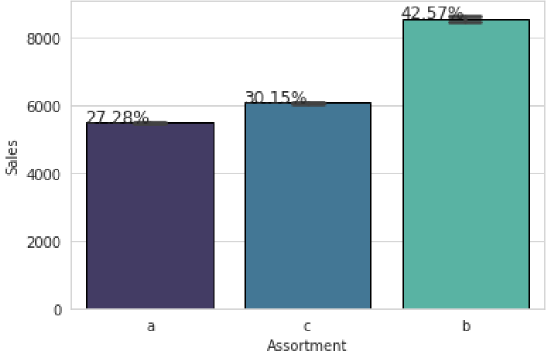
Only 17.9% of the sales are affected by the School Holiday. Rest are unaffected.



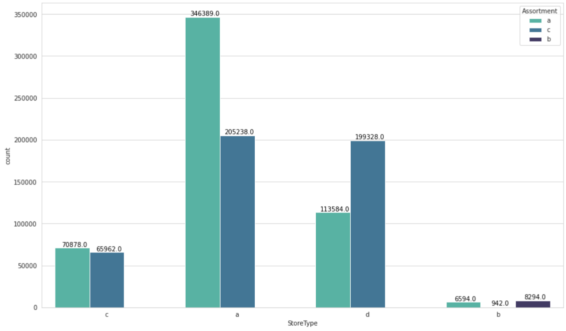
Different store types and their share in the sales. Store type B have more than 37% sales share



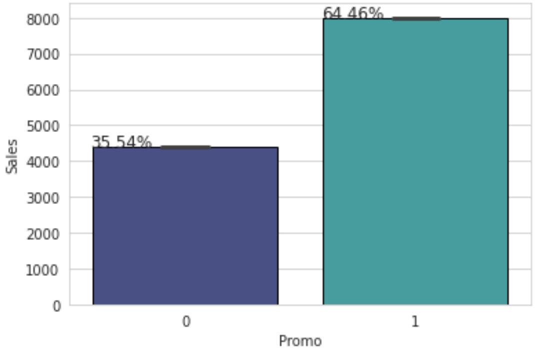
Different store types and their Customer share.Store type A have more than 50% Customer share.



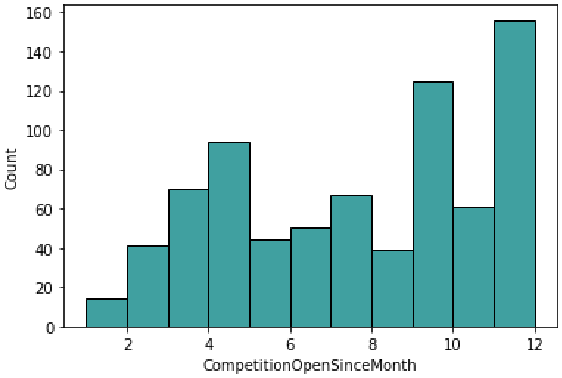
Different Assortments and their respective sales.Assortment Type B have the highest average sales , which is the major factor behind store type B's highest average sales as its a store type B exclusive.



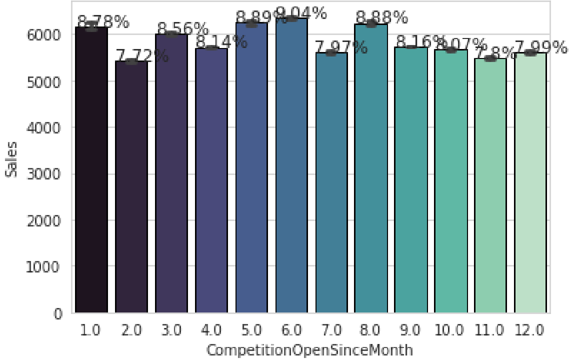
Assortment distribution across different store types.



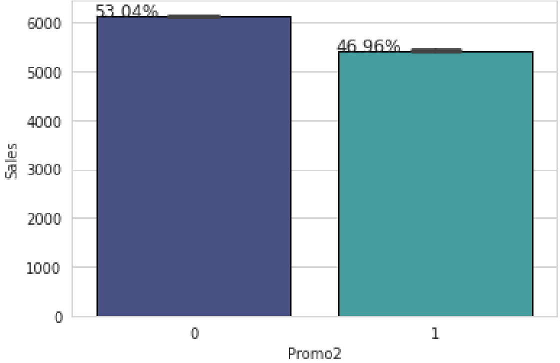
Promotion has a positive effect on Sales indicating high sales for stores with Promo=1.Sales are increased by nearly 100% after promo.



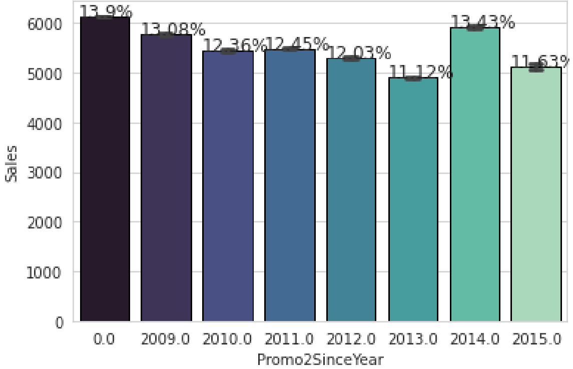
Here the above plot shows how many months have passed since the competition store opened.



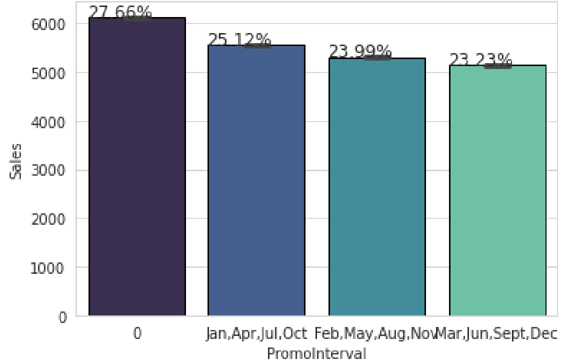
Here the above bar plot is showing the impact on sales by the opening of a Competition store.



Here the above plot shows the follow up on the first promotion.

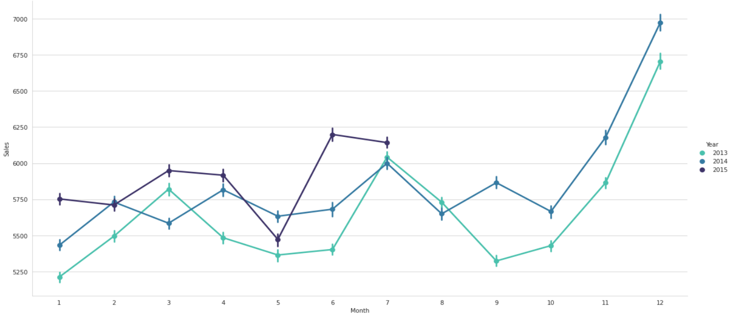


Here the above plot shows the year since the follow up promo is being run and its impact on sales.



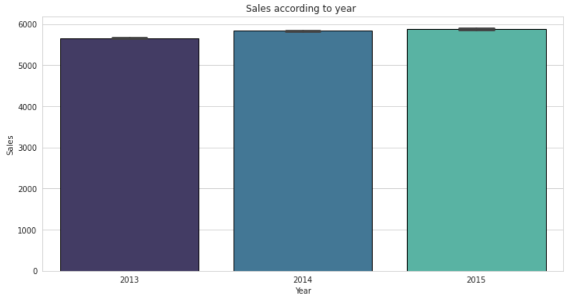
Here the plot is showing the average interval between the promo.

Continuous Insights:

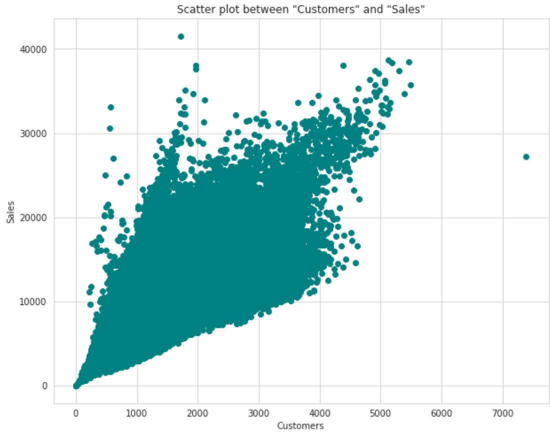


Here is a graph showing the monthly sales over time.
By the end of the year, just before the holidays, sales increase. Sales decreased from July to September in 2014, indicating that some stores had to close for renovations.

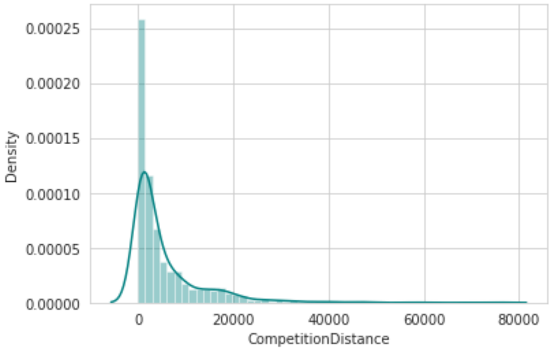
- The largest sales among the other months are in the month of December.This validates the hypothesis about this feature



Here is a bar plot showing Average sales per year from 2013 to 2015.

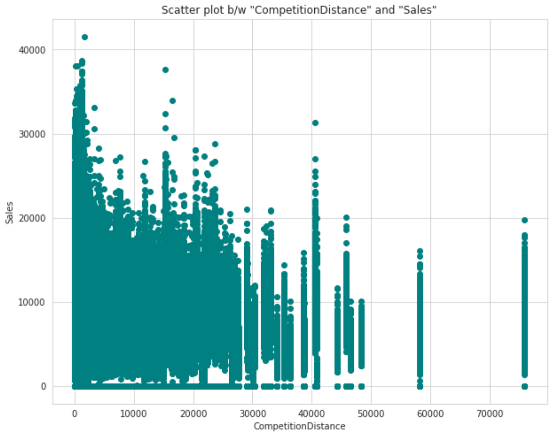


Here the Scatter plot is showing, With a few exceptions, the scatter plot of sales and customers revealed a direct, positive relationship between them.



Here the above graph is showing the Density distribution of the Competition Distance.

Most stores have competition within a distance of 0 to 10 km and have more sales than stores farther away.This validates the hypothesis about this feature.

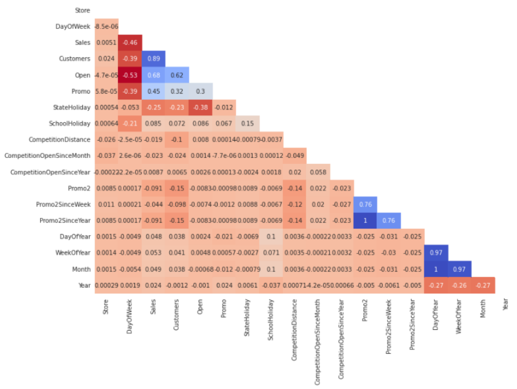


From the aforementioned scatter plot, it can be seen that the competitor stores were typically not that far apart, and the stores that were closely packed together experienced higher sales. This can be a sign of competition between populated areas and isolated areas.

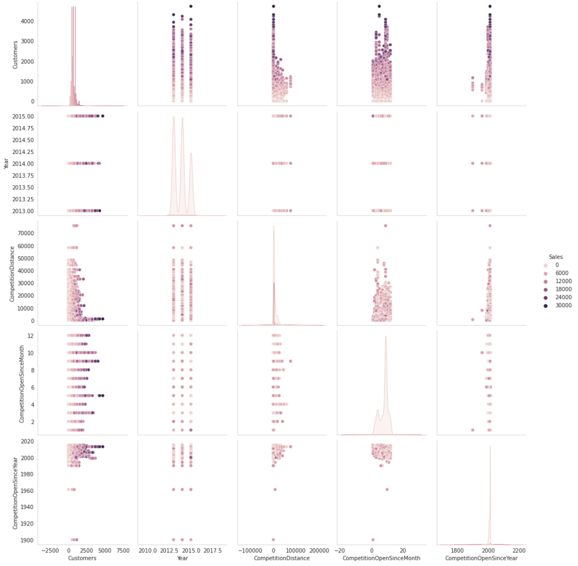
Correlation:

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 between 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]

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

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

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

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.

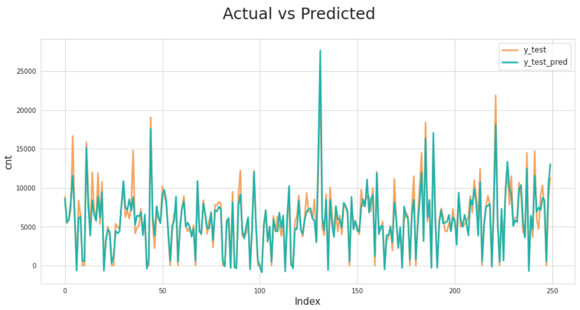
Businesses prefer the model to be interpretable in nature and decision based algorithms work better with categorical data.

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.

1.Linear Regression :

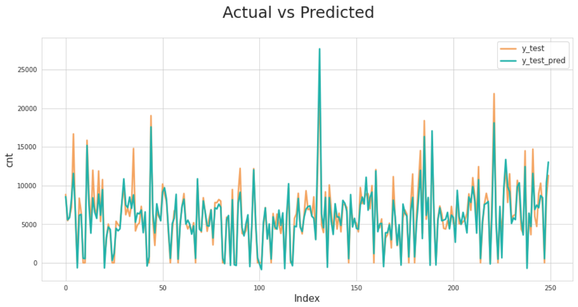
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.



The results show that a Linear Regression is performing pretty well on the validation set but it has completely overfitted the train set with a test R^2 of 0.90.

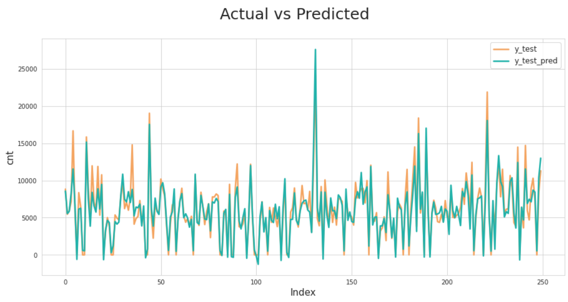
2. Ridge Regression (L2 Regularization):

Basically here, we’re going to minimize the sum of squared errors and sum of the squared coefficients (β). In the background, the coefficients (β) with a large magnitude will generate the graph peak and deep slope, to suppress this we’re using the lambda (λ) used to be called a **Penalty Factor** and help us to get a smooth surface instead of an irregular-graph. Ridge Regression is used to push the coefficients(β) value nearing **zero** in terms of magnitude. This is L2 regularization, since it's adding a penalty-equivalent to the **Square-of-the Magnitude** of coefficients.



3. Lasso Regression (L1 Regularization):

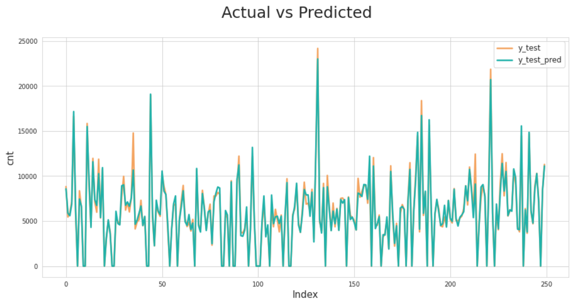
This is very similar to Ridge Regression, with little difference in Penalty Factor that coefficient is magnitude instead of squared. In which there are possibilities of many coefficients becoming zero, so that corresponding attribute/features become zero and dropped from the list, this ultimately reduces the dimensions and supports for dimensionality reduction. So, which decides that those attributes/features are not suitable as predators for predicting target value. This is L1 regularization, because of adding the **Absolute-Value** as **penalty-equivalent** to the magnitude of coefficients.



4. Decision Tree:

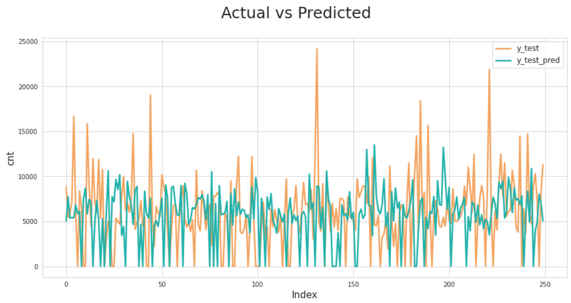
Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.



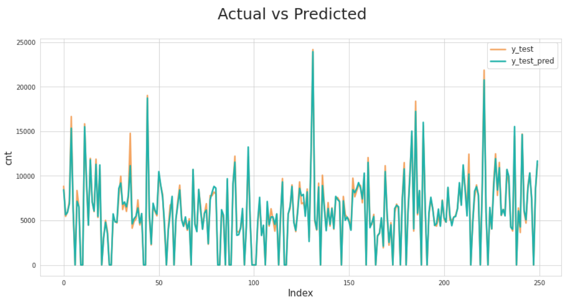
5.K-Nearest Neighbors:

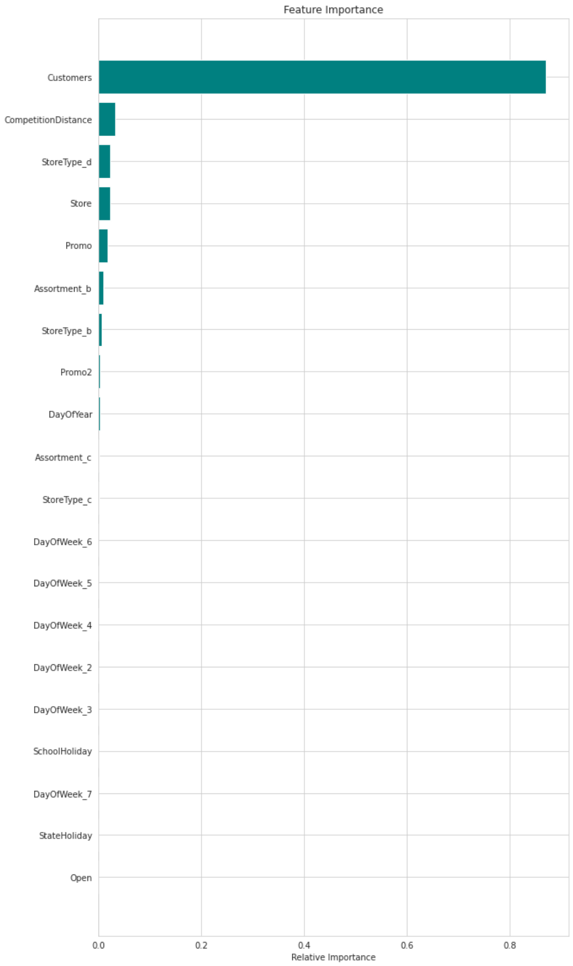
The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.



6.Random Forest Regression:

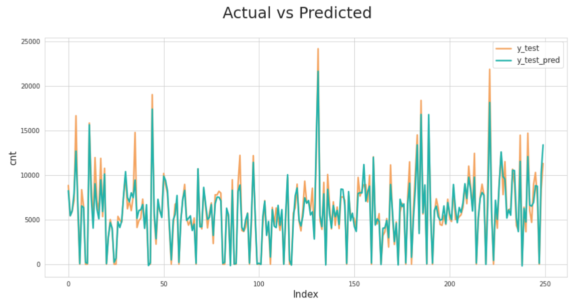
Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

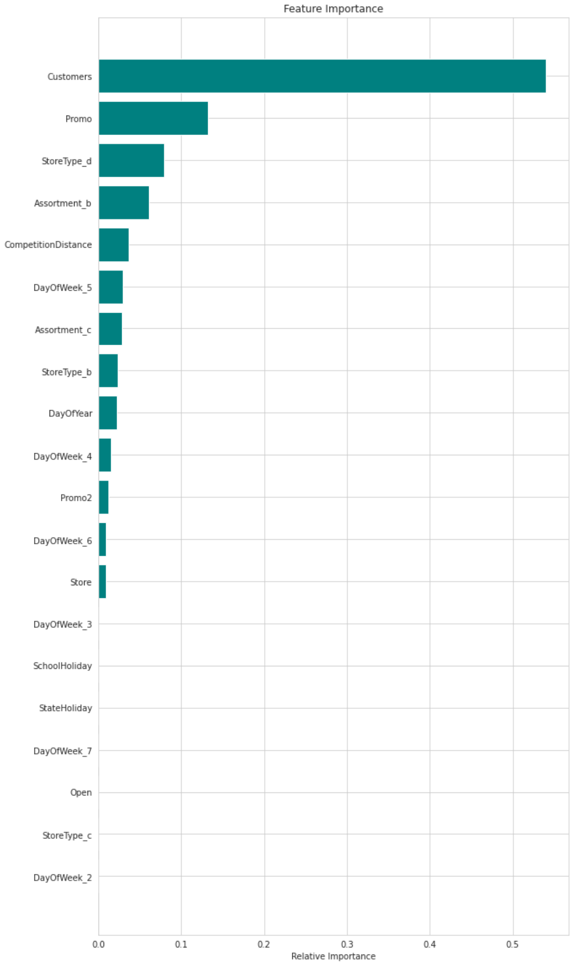




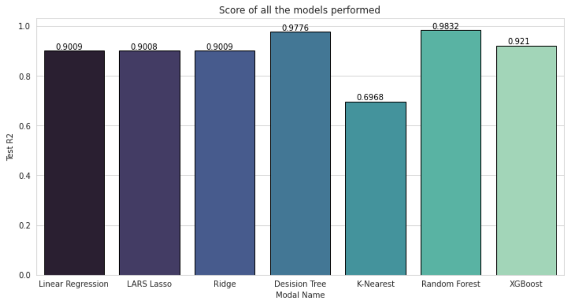
7. XGBoost:

XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values. The most common loss function in XGBoost for regression problems is reg:linear.





Model Performance and Evaluation:



Conclusion:-

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.

Following are some significant findings from the analysis:

- The largest sales among the other months are in the month of

December.This validates the hypothesis about this feature

- Due to the fact that stores are often closed on Sundays, which had the lowest sales during the week, there were more sales on Monday. This validates the hypothesis about this feature.
- Stores that participated in the promotion saw a roughly 100% increase in sales.This validates the hypothesis about this feature.
- The type of store has a significant impact on how and when stores open. Despite being scarce, store type B had the greatest average sales. The three types of assortments, especially level B, which is exclusively sold at type B stores, and the fact that the stores are open on Sundays are among the reasons. Other than for renovations or other reasons, none of the Type "b" stores ever closed.This validates the hypothesis about this feature.
- Game theory and the Nash Equilibrium are validated by the fact that most stores have competition within a distance of 0 to 10 km and have more sales than stores farther away.This validates the hypothesis about this feature.
- We can see that the majority of stores are closed on state holidays. However, it's noteworthy to notice that more stores were open during school breaks than during state holidays.
- The dataset's outliers displayed justified behavior. The anomalies either belonged to store type B or were running promotions that boosted sales.

- The XGBoost Model performs well and provides 0.92 R-Squared on the test set. All trends and patterns that could be caught by these models without overfitting were done, and the model reached its maximum level of performance.

Recommendations:-

- It is important to encourage more stores to run promotions.
- It might be possible to have more stores of type B. They have the highest average sales despite having the fewest stores.
- Because there is a seasonal component, retailers should be urged to advertise and capitalize on the holidays.

Challenges:

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

References-

1. MachineLearningMastery

2. GeeksforGeeks

3. Kaggle

4. Towards Data Science

5. Analytics Vidhya

6. Built in Data Science Blogs

7. Scikit- Learn Org