Capstone Project
Retail Sales Prediction
(Regression)

(Snehendu Jana)
Data science student
Alma Better

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 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:

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
- **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 values.

```
#null values
df.isnull().sum()
```

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

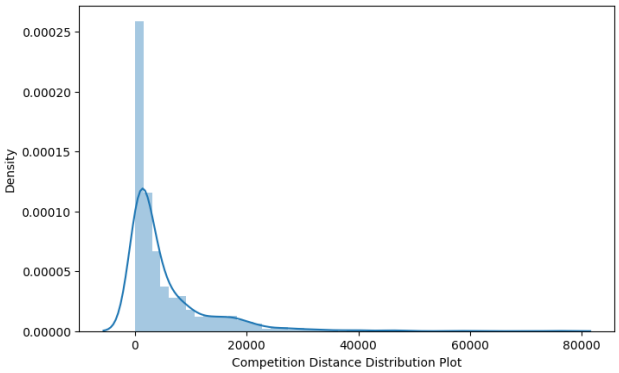
```
#null values in stores df
stores_df.isnull().sum()
```

```
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

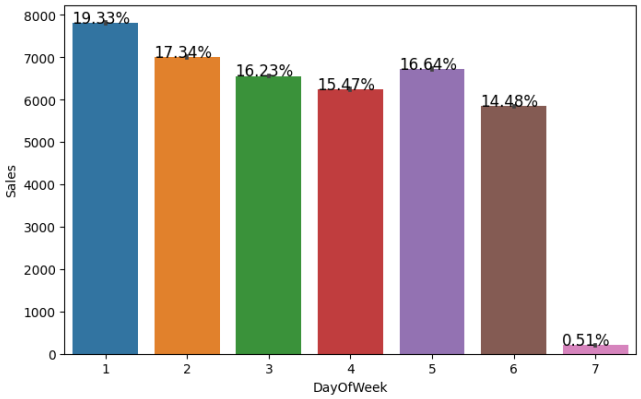
fixed, number of possible values putting a a particular category to the observation.
Non Categorical- A non categorical or continuous variable is a variable whose value is obtained by measuring, i.e., one which can take on an uncountable set of values.
Both of them are analyzed separately.
Categorical 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

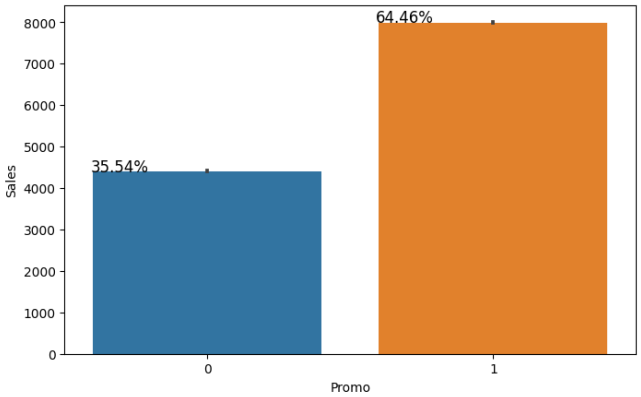
Just by observing the head of the dataset and understanding the features involved in it, the following hypotheses could be framed:

- 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.

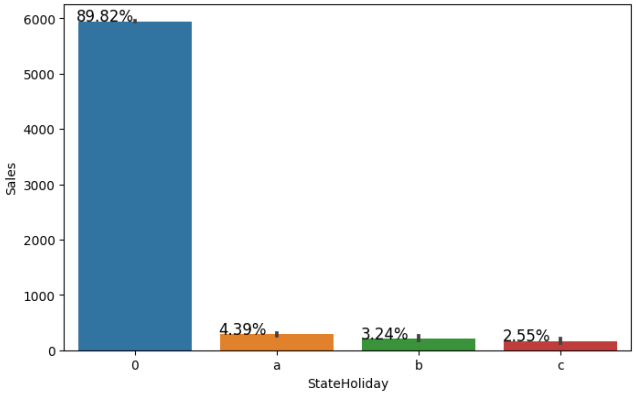
Categorical Insights:



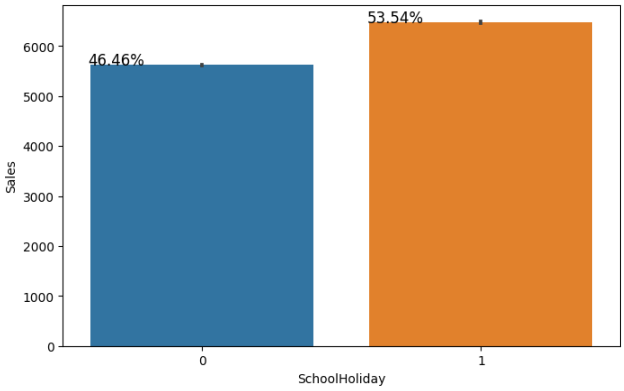
Here it can be deduced that there were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.



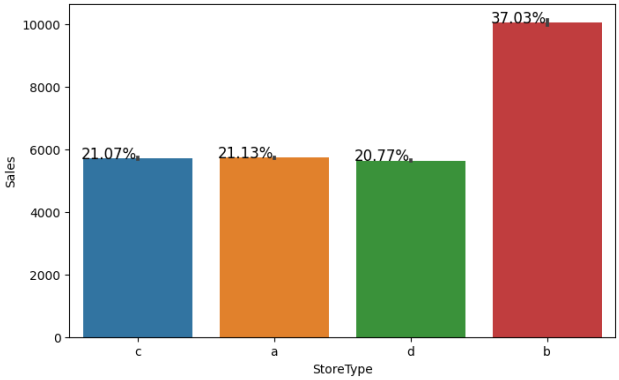
Promotion has a positive effect on Sales indicating high sales for stores with Promo=1.



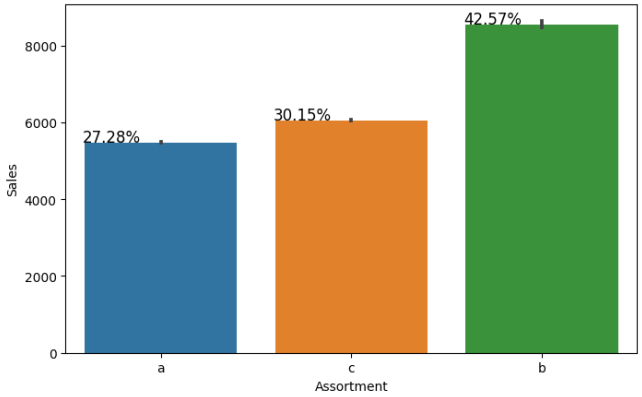
Sales were low whenever there was a State Holiday indicating only a few stores were open on these days.



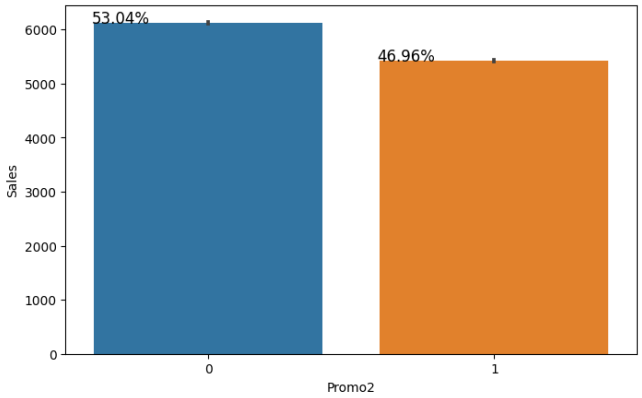
Sales were high on an average on School Holidays indicating School Holidays weren't compulsory by the law and comparatively more sales were recorded on holidays.



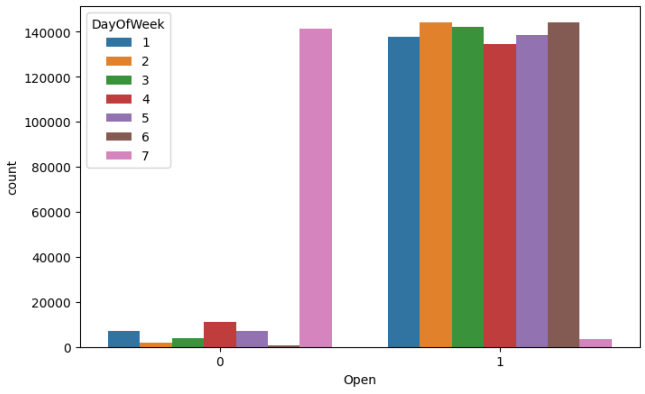
A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. The store type b has the highest sales on an average.



Assortment type b has the highest sales on an average.

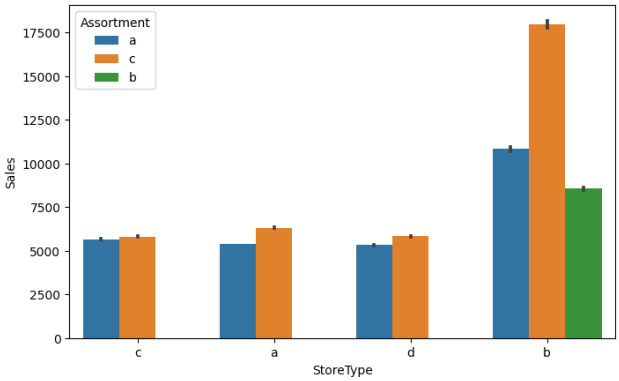


With Promo2, slightly more sales were seen without it which indicates there are many stores not participating in promo.

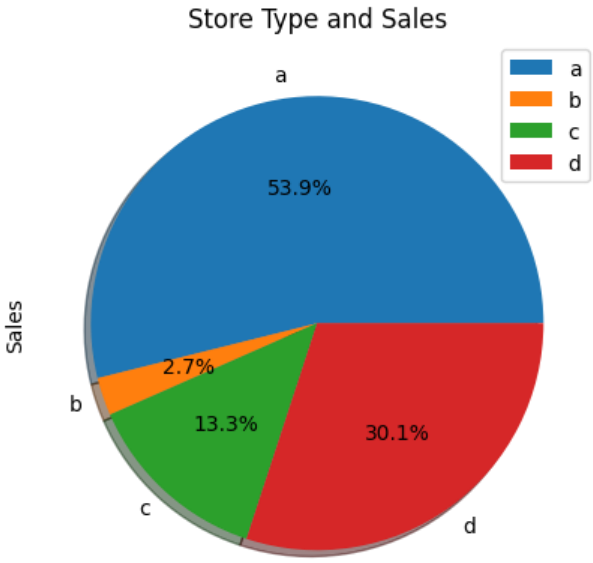


This is a count plot of open shops according to the day of the week. It's clear that the number of shops open on Sundays were very less and hence low sales. Some shops were closed on weekdays as well accounting to the stores closed due to

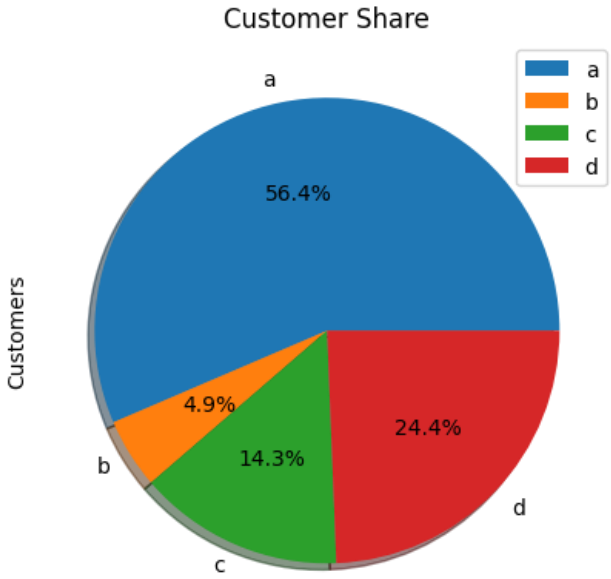
refurbishment or holidays.

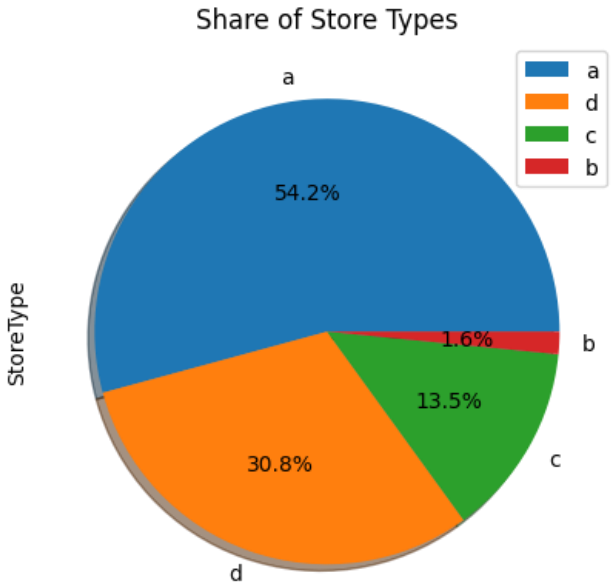


The above bar plot shows that the store types a, c and d have only assortment level a and c. On the other hand the store type b has all the three kinds of assortment strategies, a reason why average sales were high for store type b stores.



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.

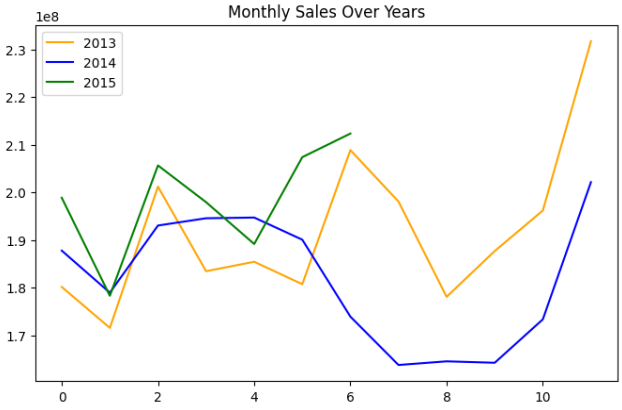




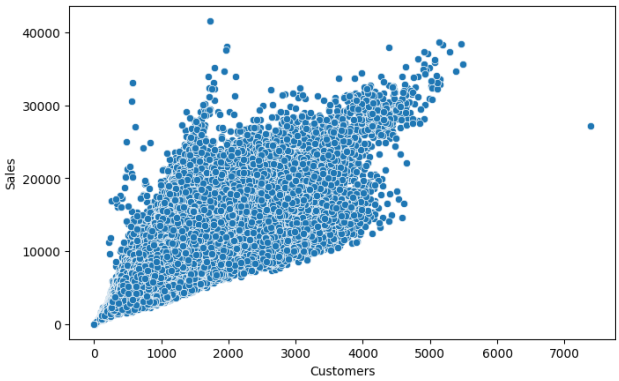
But upon further exploration it can be clearly observed that the highest sales belonged to the store type a due to the high number of type a stores in our dataset. Store type a and c had a similar kind of sales and customer share. Interesting insight to note is that store type b with highest average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

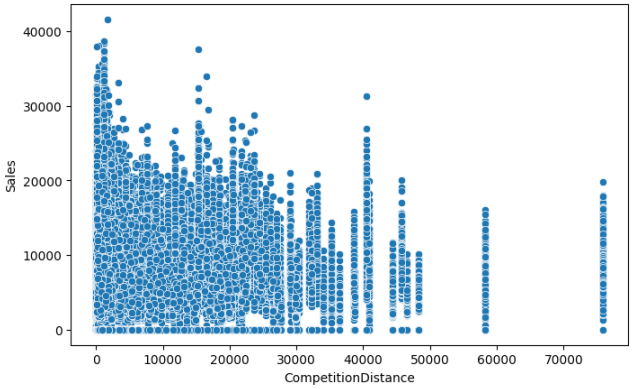
Continuous Insights:



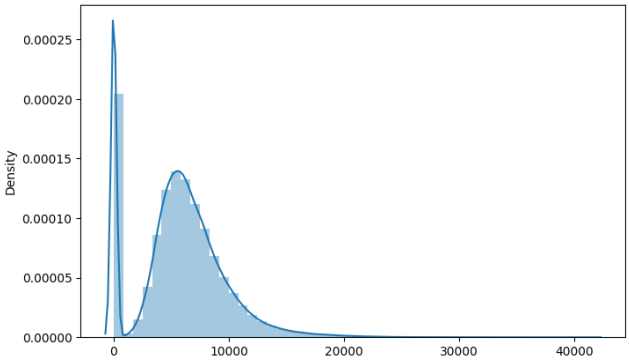
Here's a plot of Monthly Sales over the years. Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.



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.



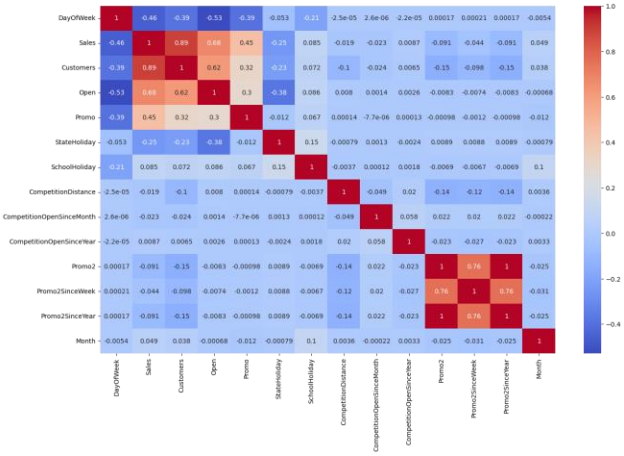
Here's a distribution plot of the Sales column. The drop in sales indicates the 0 sales accounting to the stores temporarily closed due to refurbishment.

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:

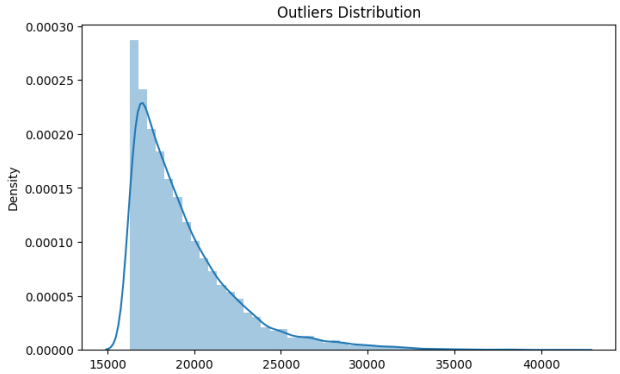
- Some stores were closed due to refurbishment and some on account of week off or holidays. Those stores on those dates generated zero sales and hence removing the rows was important to avoid confusion by the algorithms and then removing the feature altogether because it wasn't providing any value in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

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

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.

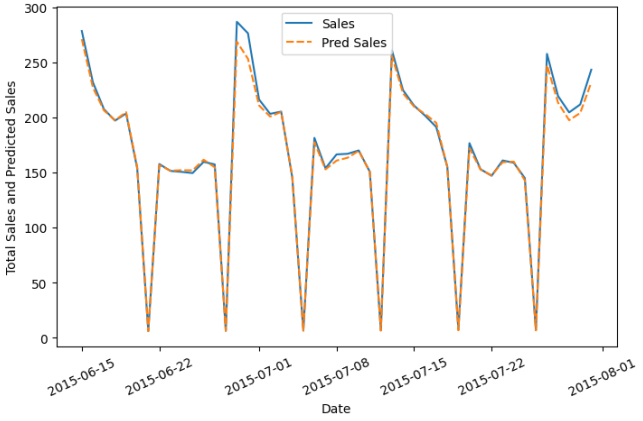
Train-Test Split:

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.
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.



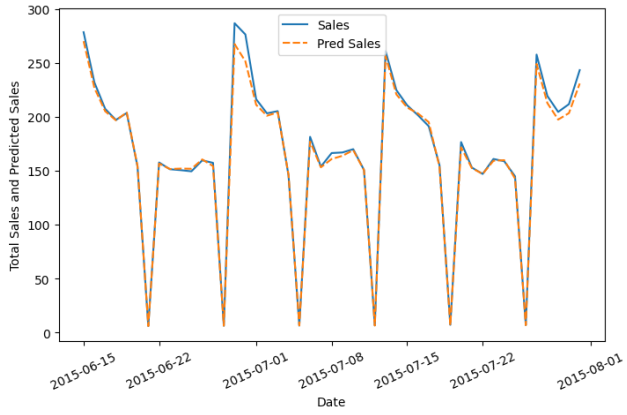
	Model_Name	Train_MAE	Train_MSE	Train_RMSE	Train_R2	Train_Adj_R2	Test_MAE	Test_MSE	Test_RMSE	Test_R2	Test_Adj_R2
0	Decision Tree Regressor	0.00000	0.000000	0.000047	1.000000	1.000000	0.014203	0.000466	0.021580	0.915750	0.915700

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.



	Model_Name	Train_MAE	Train_MSE	Train_RMSE	Train_R2	Train_Adj_R2	Test_MAE	Test_MSE	Test_RMSE	Test_R2	Test_Adj_R2
1	Random Forest Regressor	0.00304	0.000022	0.004640	0.996143	0.996143	0.010328	0.000245	0.015653	0.955673	0.955647

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

Random Forest Hyperparameters:

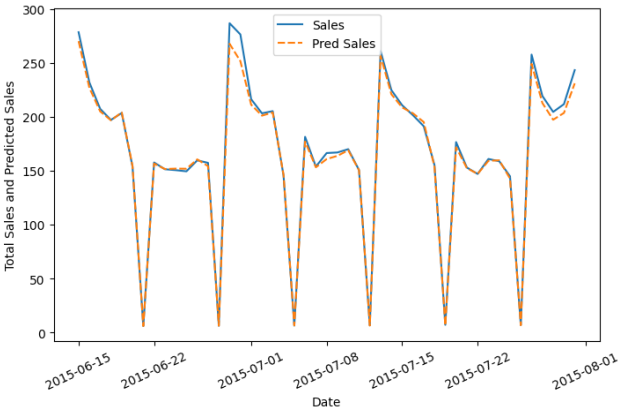
- max_depth- The max_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node
- min_sample_split- a parameter that tells the decision tree in a random forest the minimum required number of observations in any given node in order to split it.
  The default 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.

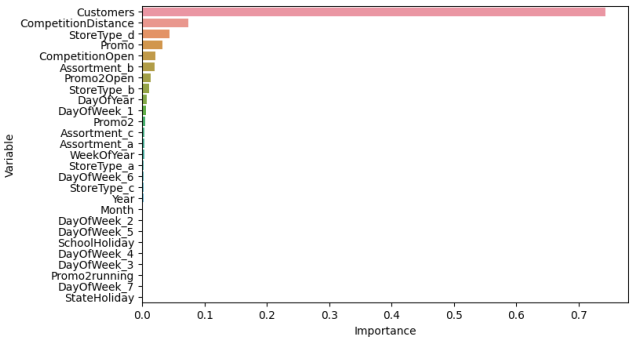
Random Forest Hyperparameter Tuned Model :

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.



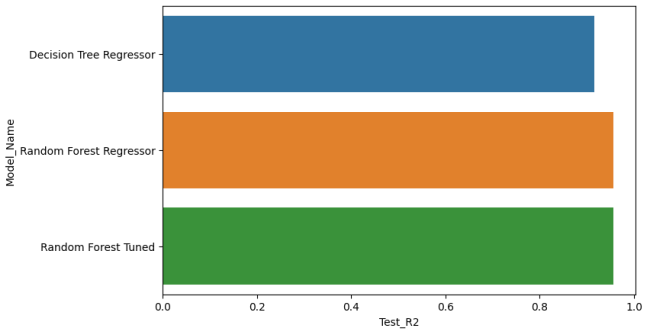
Model_Name	Train_MAE	Train_MSE	Train_RMSE	Train_R2	Train_Adj_R2	Test_MAE	Test_MSE	Test_RMSE	Test_R2	Test_Adj_R2
Random Forest Tuned	0.00304	0.000021	0.004622	0.996173	0.996173	0.010342	0.000244	0.015617	0.955878	0.955852

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:

		Sales	Pred_Sales
Date	Store		
	1	5518.0	5444.30
	2	8106.0	8087.52
2015-06-15	3	10818.0	11095.28
	4	12398.0	11685.72
	5	7808.0	7555.99

Conclusion and Recommendations:

Conclusion:

The main objective of sales forecasting is to paint an accurate picture of expected sales. Sales teams aim to either hit their expected target or exceed it.

When the sales forecast is accurate, operations go smoothly and future planning for the company's growth is done efficiently.

Upon having this analysis it can be established that given the dataset, the model developed is able to explain 95.5878 % of the variations and is able to predict the sales values in a good range.

Some important insights to draw from the analysis includes:

- There were more sales on Monday, probably because shops generally remain

closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.

- The positive effect of promotion on Customers and Sales is observable.
- Most stores have competition distance within the range of 0 to 10 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 behavior. The outliers were either of store type b or had promotion going on which increased sales.
- Random Forest Tuned Model gave the best results and only 0.021% improvement was seen from the basic random forest model which indicates that all the trends and patterns that could be captured by these models without overfitting were done and maximum level of performance achievable by the model was achieved.

Recommendations:

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

Challenges:

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

References:

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