

Capstone Project 2

Retail Sales Prediction

Individual Project:

Name: Mayank Ghai

Email: mayankghai1195@gmail.com

AImaBetter



Content

- **Problem Statement**
- **Data Pipeline**
- **Retail Sales Prediction**
- **Data Summary**
- **Approach**
- **Outlier Detection**
- **Exploratory Data Analysis**
- **Model performed**
- **Model Performance and Evaluation**
- **Store-wise Sales Predictions**
- **Conclusion and Recommendations**
- **Challenge**



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.



Data Pipeline

- **Exploratory Data Analysis (EDA):** In this part, we have done some EDA on the features to see the trend.
- **Data Processing:** In this part, we went through each attribute and processed the dataset according to our needs.
- **Model Creation:** Finally in this part we created the various models. These various models are being analysed and we tried to study various models so as to get the best performing model for our project.



Retail Sales Prediction

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 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 different machine learning algorithms.



Data Summary

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



- **Id** - an Id that represents a (Store, Date) tuple 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 started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store.



Approach

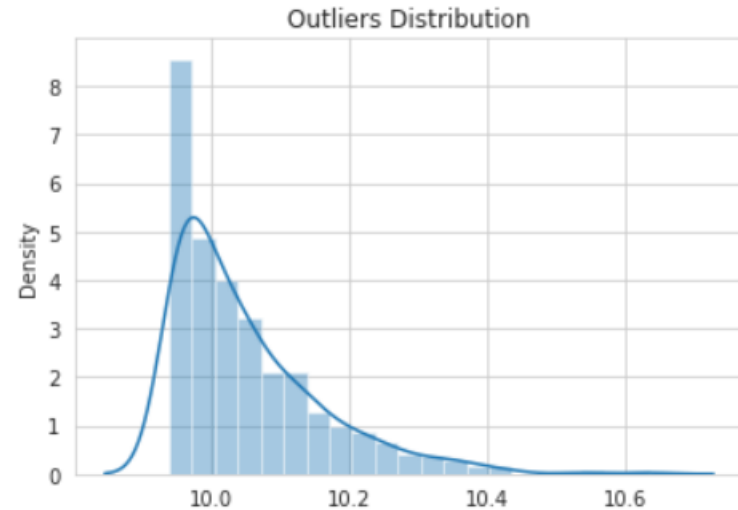
The following approach was followed in the completion of the project:

- **Business Problem**
- **Data Collection and Preprocessing**
 - Data Cleaning
 - Missing Data Handling
 - Merging the Datasets
- **Data Manipulation**
 - Feature Engineering
 - Outlier Detection and Treatment
- **Exploratory Data Analysis**
 - Hypotheses
 - Categorical Features
 - Continuous Features
 - EDA Conclusion and Validating Hypotheses
- **Modeling**
 - Feature Scaling
 - Categorical Data Encoding
 - Train Test Split
 - Baseline Model
 - Hyperparameter Tuning
 - Feature Importance
- **Model Performance and Evaluation**
 - Visualizing Model Performances
 - Comparison among the models
- **Store-wise Sales Predictions**
- **Conclusion and Recommendations**



Outlier Detection

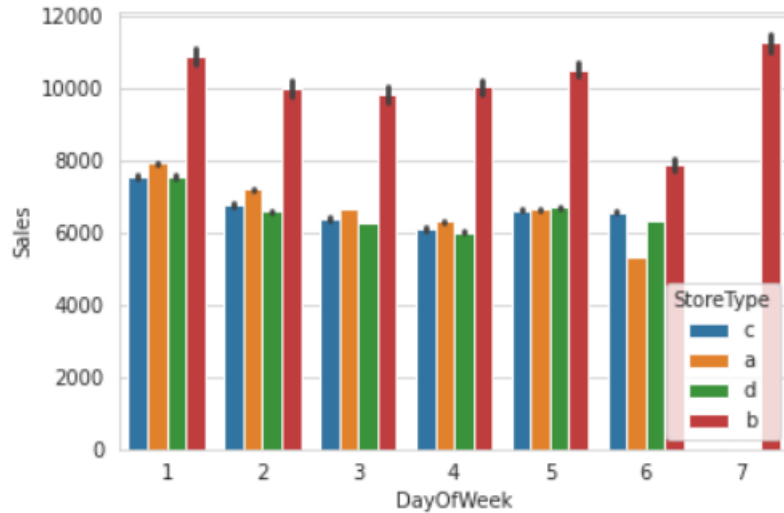
- In statistics, an outlier is a data point that differs significantly from other observations. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.
- Z-score is a statistical measure that tells you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



		DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
Date	Store															
2013-01-06	85	7	10509	1509	1	0	0	0	b	a	1870.0	10.0	2011.0	0	NaN	NaN
	259	7	7926	1686	1	0	0	1	b	b	210.0	9.0	2013.0	0	NaN	NaN
	262	7	23240	3479	1	0	0	0	b	a	1180.0	5.0	2013.0	0	NaN	NaN
	274	7	3802	932	1	0	0	1	b	b	3640.0	9.0	2013.0	1	10.0	2013.0
	310	7	2334	193	1	0	0	0	a	c	2290.0	9.0	2013.0	1	10.0	2014.0
...
2015-07-26	948	7	12040	2346	1	0	0	0	b	b	1430.0	9.0	2013.0	0	NaN	NaN
	1045	7	5968	832	1	0	0	0	a	c	26990.0	12.0	2013.0	0	NaN	NaN
	1081	7	5766	875	1	0	0	0	b	a	400.0	3.0	2006.0	0	NaN	NaN
	1097	7	13307	2710	1	0	0	0	b	b	720.0	3.0	2002.0	0	NaN	NaN
	1099	7	5683	962	1	0	0	0	a	c	200.0	4.0	2013.0	1	14.0	2013.0

3593 rows x 20 columns





- It can be well established that the outliers are showing this behaviour for the stores with promotion = 1 and store type B. It would not be wise to treat them because the reasons behind this behaviour seems fair and important from the business point of view.
- 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 datasets tree based machine learning algorithms are used which are robust to outlier effect.
- Being open 24*7 along with all kind of services available is probably the reason why it has the highest average sales than any other store.



Exploratory Data Analysis



Hypotheses

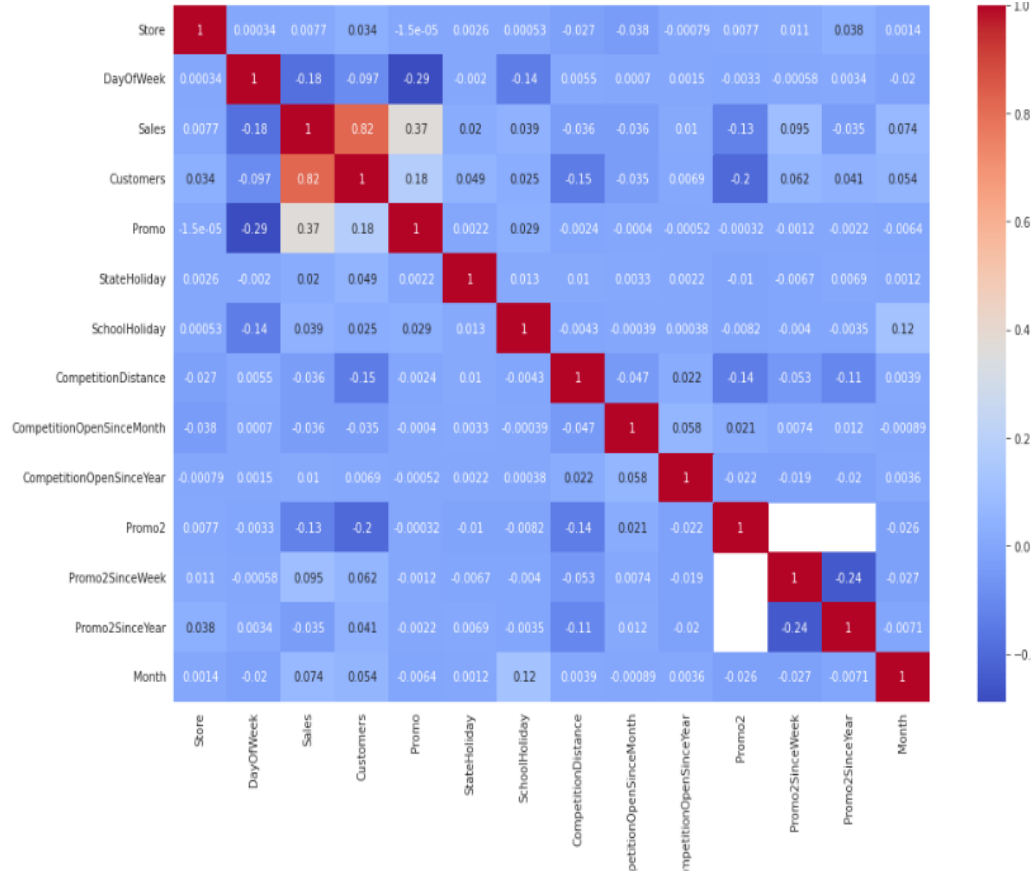
Just by observing the head of the dataset and understanding the features involved in it, the following hypotheses could be framed:

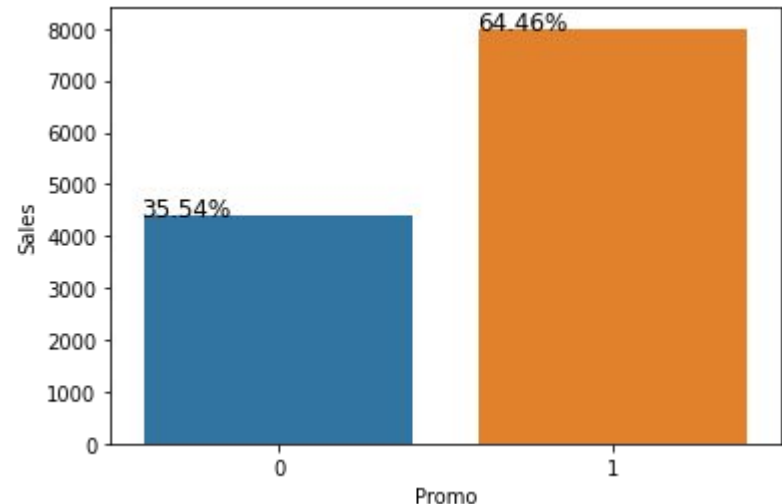
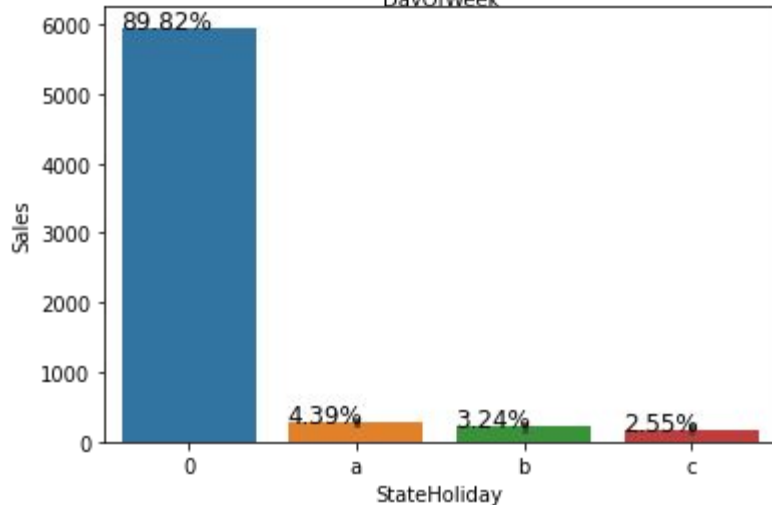
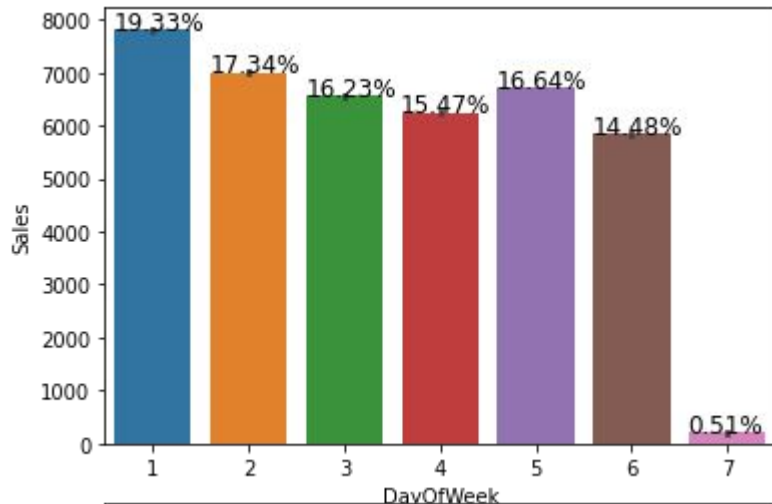
- There's a feature called "DayOfWeek" with the values 1-7 denoting each day of the week. There would be a week off probably Sunday when the stores would be closed and we would get low overall sales.
- Customers would have a positive correlation with Sales.
- The Store type and Assortment strategy involved 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 are closed due to refurbishment, those would generate 0 revenue for that time period.
- There would be some seasonality involved in the sales pattern, probably before holidays sales would be high.



EDA - Feature Correlation

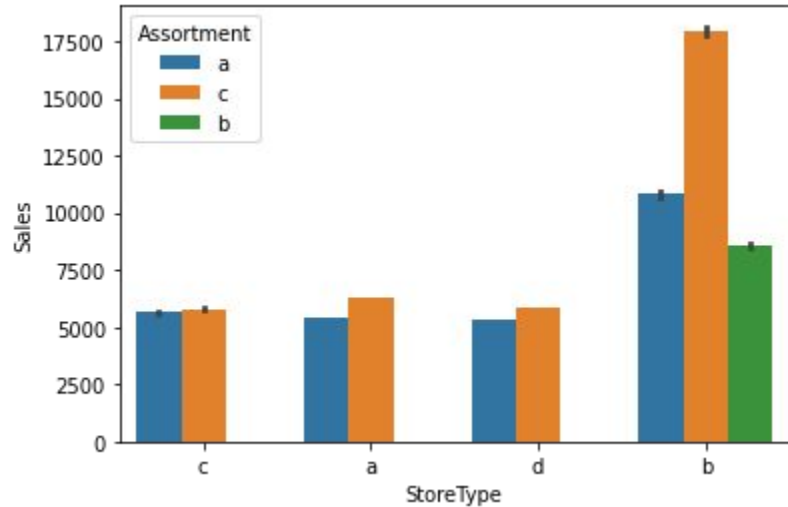
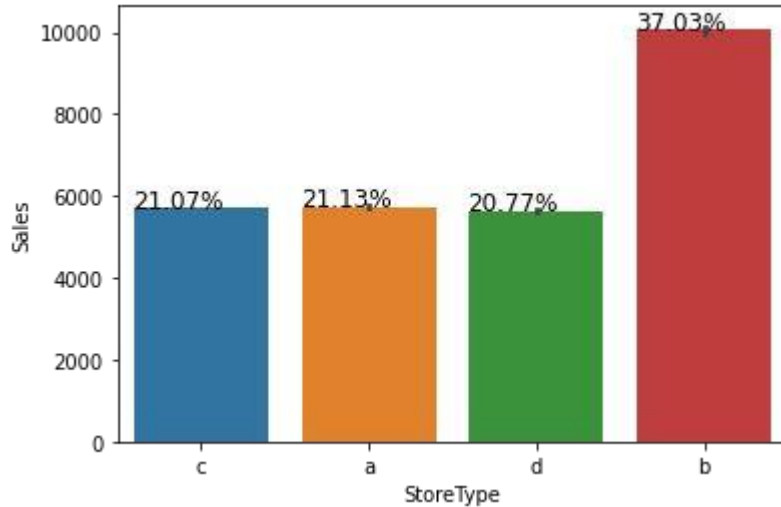
- Day of the week has a negative correlation indicating low sales as the weekends, and promo, customers and open has a positive correlation.
- State Holiday has a negative correlation suggesting that stores are mostly closed on state holidays indicating low sales.
- CompetitionDistance showing a negative correlation suggests that as the distance increases sales reduce.
- 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.





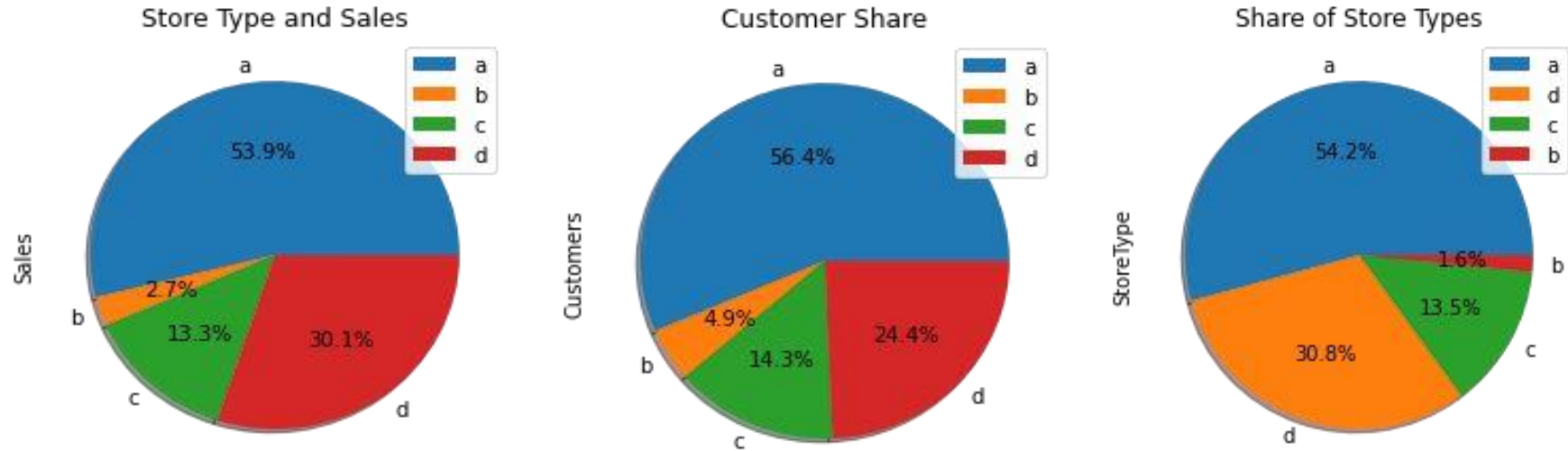
- There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week.
- Promo leads to more sales.
- Normally all stores, with few exceptions state holidays. Lowest of Sales were holidays especially on Christmas.





- A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. Here, it can be seen that on an average Store type B had the highest sales. There has to be something different about this store type.
- Next it can be seen 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.

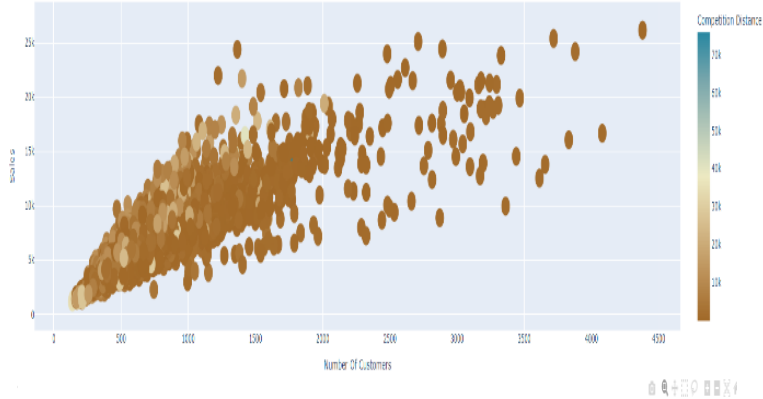




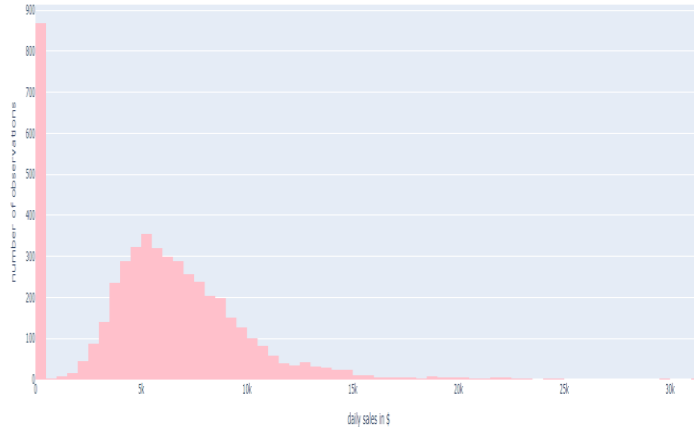
- Upon further exploration it can be clearly observed that the highest sales belonged to the store type 'a' due to the high number of type a stores in our dataset. Store type a and c had a similar kind of sales and customer share.
- Based on the above findings it seems that there are quite a lot of opportunities in store type a & c 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 customer share, because the majority of the stores were of these kinds, they had the best overall revenue. On the other hand, store type b were very few in number and even then they had better sales than others.



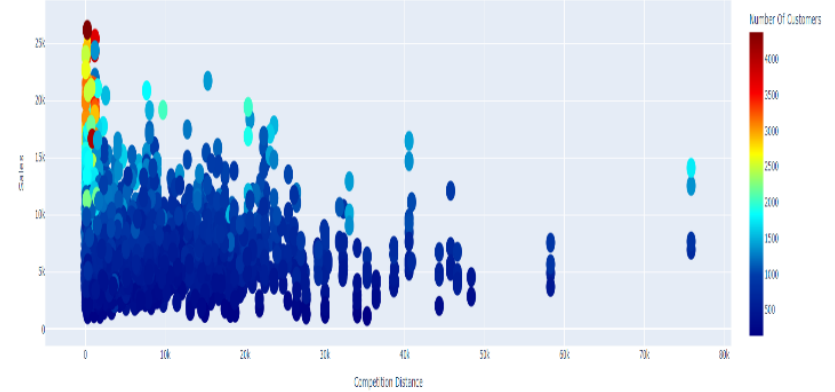
Sales vs. Number Of Customers



Sales Distribution

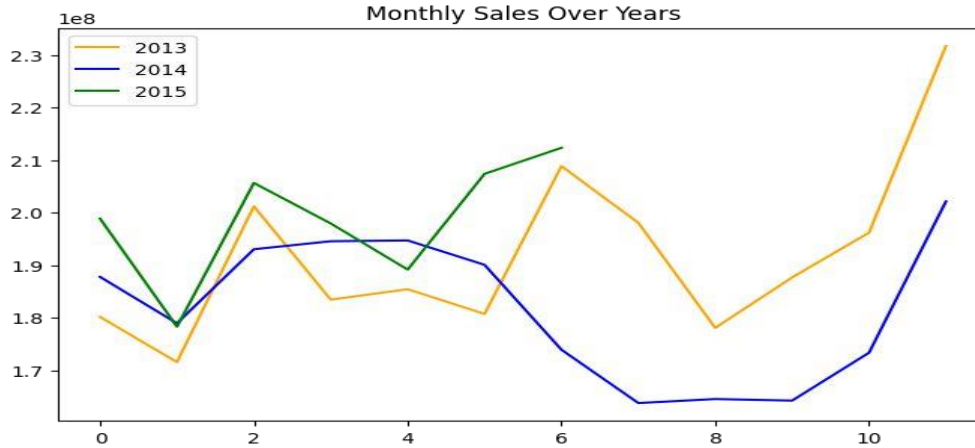


Sales vs. Competition Distance

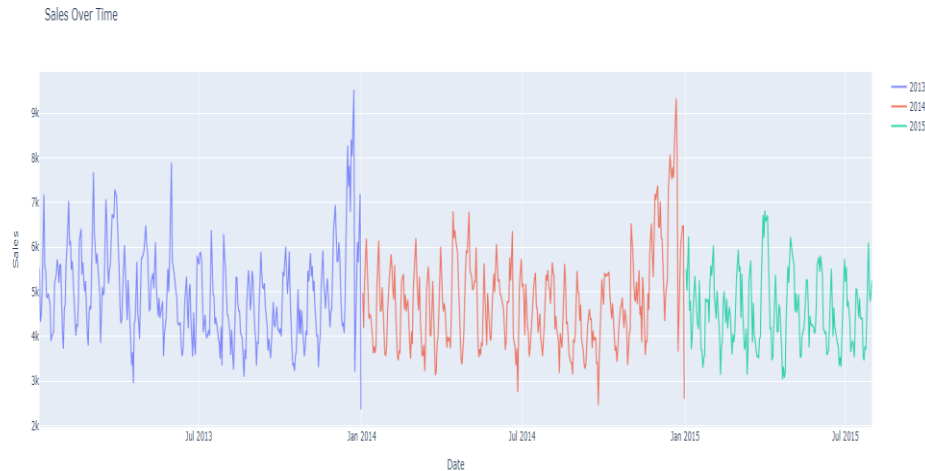


- It's pretty obvious that there is going to be a positive correlation between customers and sales. There are a few outliers.
- Most stores have competition distance within the range of 0 to 10 kms and had more sales the further away.
- The drop in sales indicates the 0 sales the stores temporarily closed due to

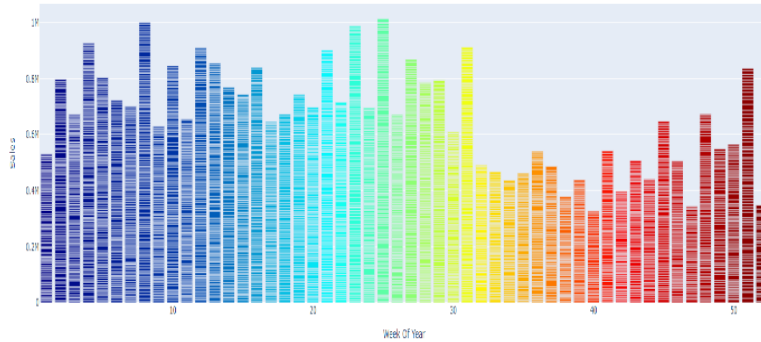




- Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.

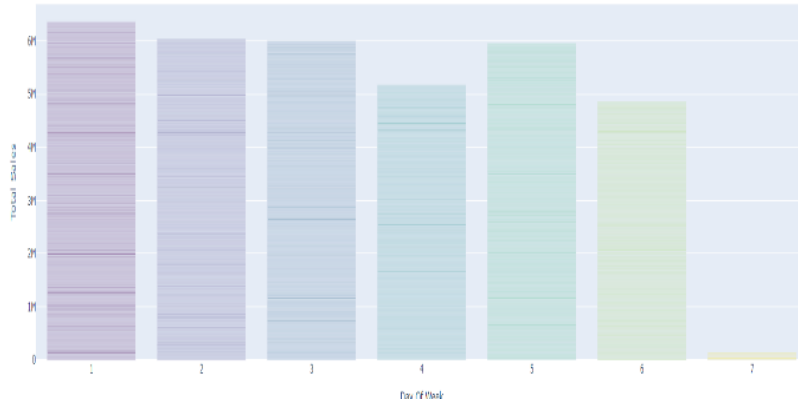


Sales Over Week Of Year



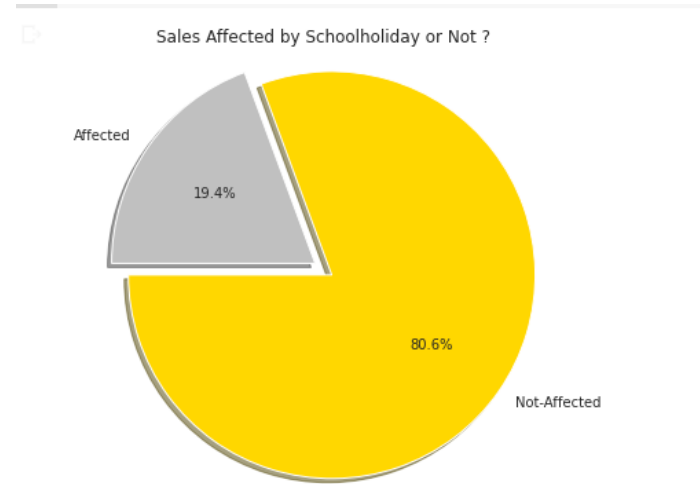
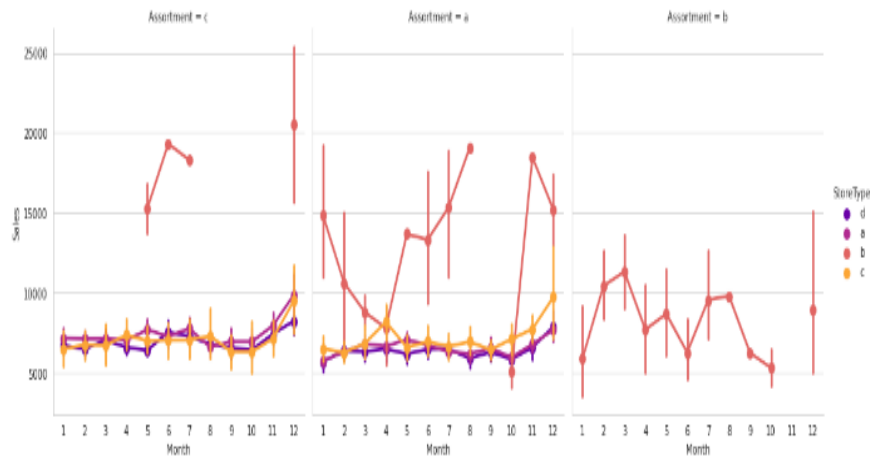
- The graph gives us the knowledge of the weeks when the number of sales has been the most and when it is recorded as the lowest.

Sales Over Days Of A Week

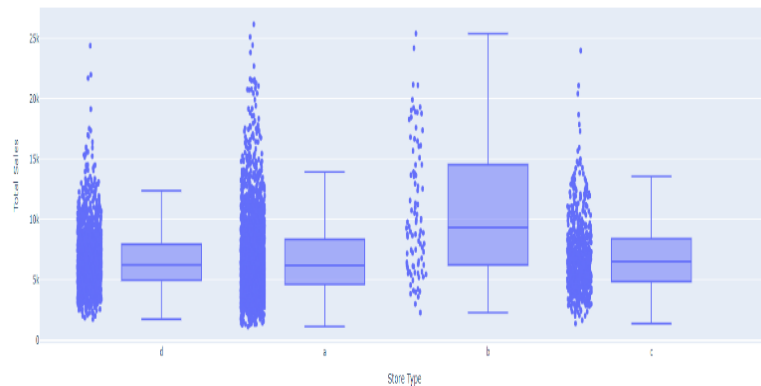


- The graph shows the total sales over the week that the store had by combining the sales that happened on the days of the week over the whole year.





Sales By Store Type



- **Earlier, it was observed that only store type b had all three kinds of assortment levels and the rest of the store types had two of them. It seems that in some b type stores the products were different as compared to others because the revenue per store is significantly more than the other**
- It can be seen that the Sales are not affected



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



Model's Performed

- **Linear Regression**
- **Bayesian Ridge Regression**
- **LARS Lasso Regression**
- **Decision Tree Regression**
- **Random Forest Regression**
- **K-Nearest Neighbours Regression**
- **Random Forest Regression Tuned**




Model's Evaluation Matrices

	Model_Name	Regression Model Score	Sample Test Score	Training RMSE	Testing RMSE	Training MAPE	Testing MAPE	R2_train	R2_test	Adj_r2_train	Adj_r2_test	
0	LinearRegression	0.668434	0.664806	0.234464	0.236513	1.958963	1.975089	0.668434	0.664806	0.668197	0.664246	
1	BayesianRidge	0.668427	0.664814	0.234466	0.236510	1.958863	1.974963	0.668427	0.664814	0.668190	0.664255	
2	LassoLars	0.468717	0.468717	0.296793	0.299353	2.634731	2.651230	0.468717	0.463026	0.468337	0.462130	
3	DecisionTreeRegressor	0.824154	0.750278	0.170748	0.204143	1.406284	1.678928	0.824154	0.750278	0.824054	0.749944	
4	RandomForestRegressor	0.975291	0.831089	0.170748	0.204143	1.406284	1.678928	0.824154	0.750278	0.824054	0.749944	
5	KNeighborsRegressor	0.581671	0.477215	0.263360	0.295371	2.361319	2.641302	0.581671	0.477215	0.581372	0.476342	
6	RandomForestRegressorTuned	0.976877	0.848017	0.061917	0.159259	0.519827	1.321998	0.976877	0.848017	0.976864	0.847814	

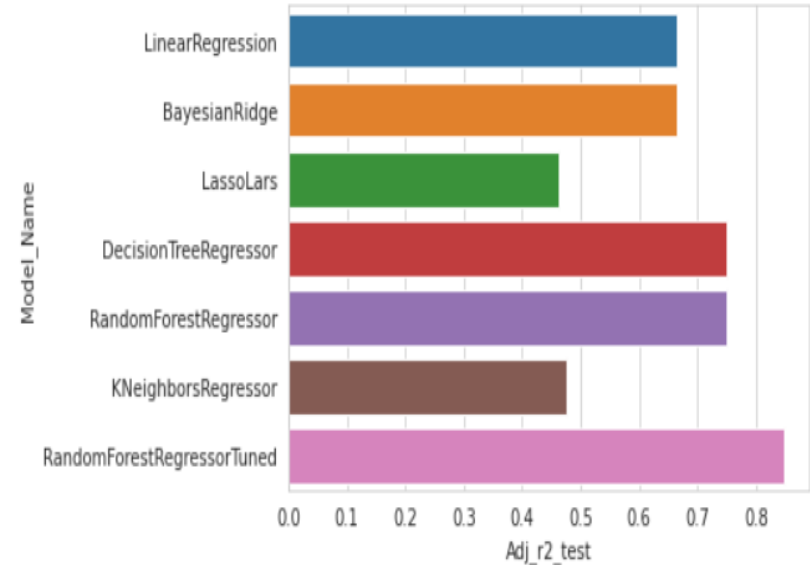
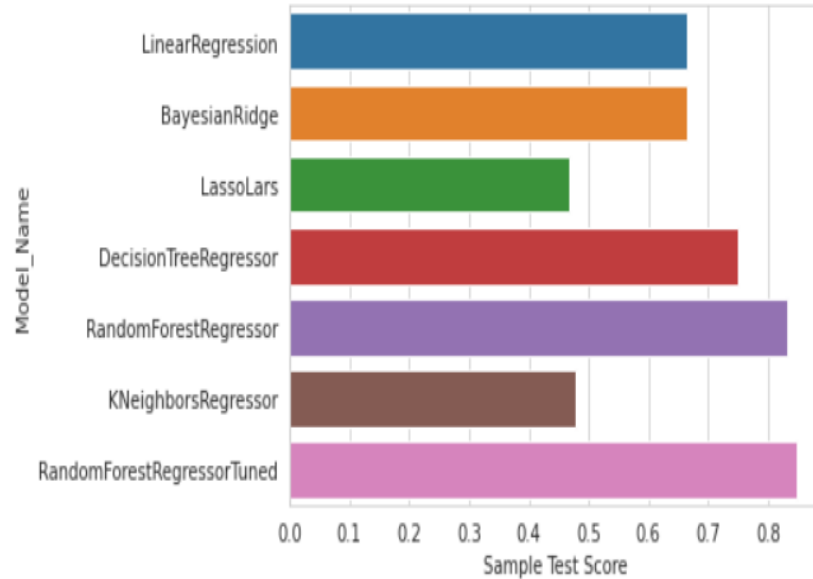


Sample Test Score, R2 Score, and Adjusted R2 of Model's Performed

	Model_Name	Sample Test Score	R2_test	Adj_r2_test	
0	LinearRegression	0.664806	0.664806	0.664246	
1	BayesianRidge	0.664814	0.664814	0.664255	
2	LassoLars	0.468717	0.463026	0.462130	
3	DecisionTreeRegressor	0.750278	0.750278	0.749944	
4	RandomForestRegressor	0.831089	0.750278	0.749944	
5	KNeighborsRegressor	0.477215	0.477215	0.476342	
6	RandomForestRegressorTuned	0.848017	0.848017	0.847814	



Sample Test Score and Adjusted R2 of Model's Performed



- Since the R^2 Score and the Adjusted r^2 Score is similar in two of the three cases and hence we have to shift to the Sample Test Score.
- The sample model scores of **Decision Tree** is **0.7502775387784837**, **Random Forest** is **0.831089330659996** and **Tuned Random Forest** is **0.8480172625761316**
- Improvement of **10.771 %** was seen in **Random Forest** against **Decision Tree**
- The Sample Test Score was seen in tuned Random Forest model with the value **0.848** which was only **2.037 %** improvement over a simple random forest model.

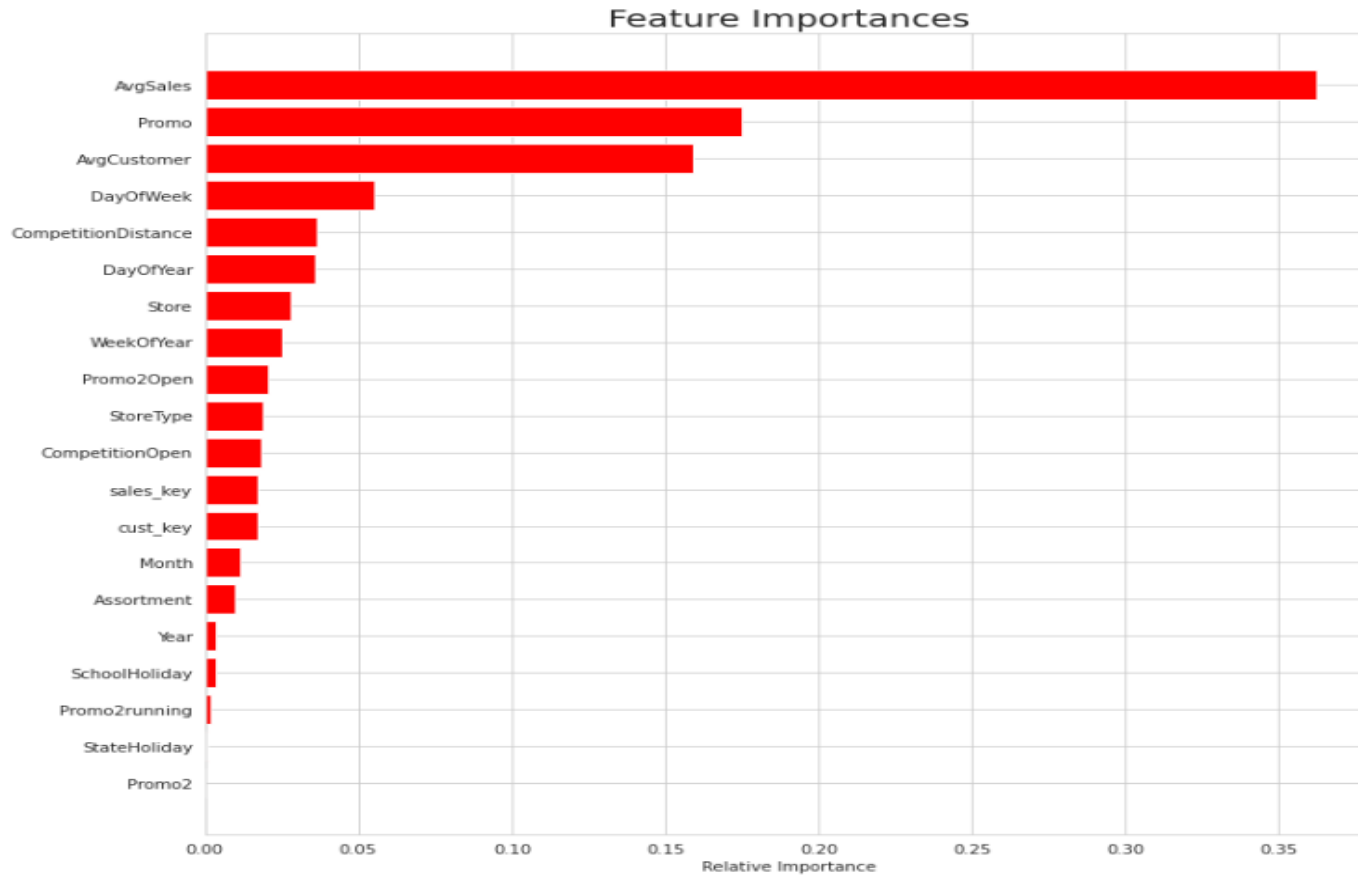


Model Validation & Selection

- **Observation 1:** As seen in the Model Evaluation Matrices table, Linear Regression, KNN is not giving great results.
- **Observation 2:** Decision Tree & Random Forest have the same value of R^2 and Adjusted R^2 .
- **Observation 3:** Random forest & Tuned Random Forest have performed equally good in terms of Sample Model Score, R^2 Score and Adjusted R^2 Score.



Feature Importance



Model Performance and Evaluation


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. Hence, a simple decision tree was used as a baseline model.
- As seen in the Model Evaluation Matrices table, Linear Regression, KNN is not giving great results.
- The sample test scores of **Decision Tree are 0.750, Random Forest is 0.831** , and **Tuned Random Forest is 0.84**.
- Improvement of **10.771 %** was seen in **Random Forest against Decision Tree**.
- The Sample Test Score was seen in the tuned Random Forest model with the value of **0.848** which was only **2.037 %** improved from a simple random forest model.



Store wise Sales Predictions

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



		Sales	Pred_Sales
Date	Store		
2014-04-07	393	5261	4756.695842
2014-10-08	576	7693	7229.458791
2014-09-26	751	2459	3550.183462
2013-11-02	795	5237	4046.713663
2013-08-09	53	5221	5725.194103



Conclusion and Recommendations:

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.

Some important conclusions drawn from the analysis are as follows:

- The positive effect of promotion on Customers and Sales.
- Most stores have competition distance within the range of 0 to 10 km and had more sales than stores far away probably indicating competition in busy locations vs remote locations.
- Store type B though being few in number had the highest sales average. The reasons include all three kinds of assortments especially assortment level b which is only available at type b stores and is open on Sundays as well.
- The outliers in the dataset showed justifiable behavior. The outliers were either of store type b or had a promotion going on that increased sales.

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

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



**THANK
YOU**

