Capstone Project 2

Retail Sales Prediction

Group Project

Team Members:

Annayan Bose
Prabhat Patel

Content

- Problem Statement
- Retail Sales Prediction
- Data Summary
- Approach
- Exploratory Data Analysis
- Outlier Detection
- Modeling:
  - Linear Regression
  - Baseline Model – Decision Tree
  - Random forest
  - Random Forest Hyperparameter Tuning
- Model Performance and Evaluation
- Store wise Sales Predictions
- Conclusion and Recommendations

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.

Retail Sales Prediction

Sales forecasting is the technique of predicting demand or sales of a specific product over a predetermined time frame.Businesses use sales forecasts to estimate the amount of income they will bring in over a specific period of time so they may create strong and effective business plans.

The income the firm expects to generate in the upcoming months has an impact on crucial decisions like budgets, hiring, incentives, objectives, acquisitions, and numerous other growth plans, thus it's critical that these projections be accurate for the plans to be as successful as they are intended to be.

The 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

- **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.

Approach

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

- **Business Problem**
- **Data Collection and Preprocessing**
  - **Datasets exploration**
  - **Missing Data Handling**
  - **Merging the Datasets**
- **Exploratory Data Analysis**
  - **Categorical Features**
  - **Continuous Features**
- **Data Manipulation**
  - **Outlier Detection and Treatment**
  - **Feature Engineering and Feature Selection**
  - **Categorical Data Encoding**
- **Modeling**
  - **Train Test Split**
  - **Feature Scaling**
  - **Linear Regression**
  - **Baseline Model – Decision Tree**
  - **Random Forest**
  - **Random Forest Hyperparameter Tuning**
- **Model Performance and Evaluation**
  - **Feature Importance**
  - **Visualizing Model Performances**
- **Conclusion and Recommendations**

Dataset Exploration

We have two datasets. Rossman store data is for years 2013, 2014 and 2015 with 10,17,209 observations on 9 variables. Stores data with 1115 observations on 10 variables. Some important features are:

1. Customer : - The number of customers on a given day in a store.
2. Date :- Showing dates for observations.
3. State Holiday :- Indicating a state holiday.
4. Store Type :- Differentiate between 4 different store models (a,b,c,d).
5. Assortment :- Describes an assortment level i.e a : basic, b : extra and c : extended.
6. Competition Distance : Distance in meters to the nearest competition store.
7. Promo :- Indicates whether a store is running a promo on that day
8. Promo2 :- Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating.
9. Open - an indicator for whether the store was open: 0 = closed, 1 = open.
10. Sales - the turnover for any given day (Dependent Variable).

Missing Value Treatment

Rossman Dataset – No Null Values

StoreDataset - Out of 10 columns there are missing values for the below columns:

CompetitionDistance- distance in meters to the nearest competitor store, the distribution plot would give us an idea about the distances at which generally the stores are opened and we would impute the values accordingly.

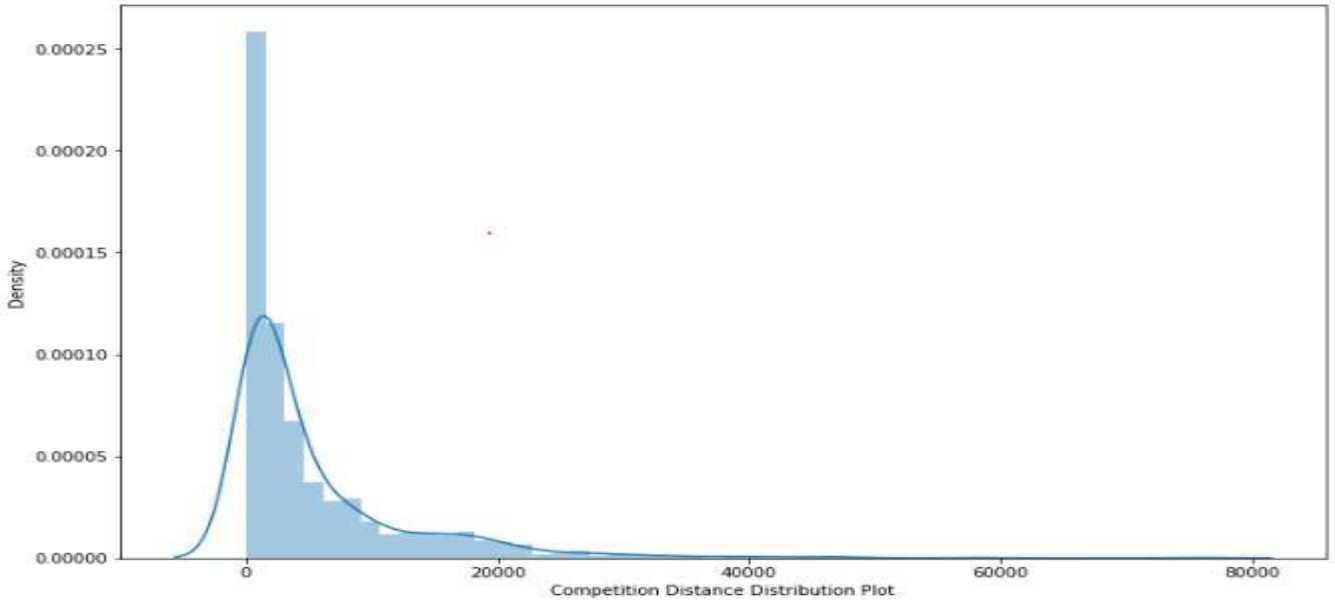
CompetitionOpenSinceMonth- gives the approximate month of the time the nearest competitor was opened, mode of the column would tell us the most occuring month.

CompetitionOpenSinceYear- gives the approximate year of the time the nearest competitor was opened, mode of the column would tell us the most occuring 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 can be replaced with 0.

Distribution of CompetitionDistance

It seems like most of the values of the CompetitionDistance are towards the left and the distribu tion is skewed on the right. Median is more robust to outlier effect. So we will fill the nulls with median.



Now we don’t have any null values in both the datasets.

Exploratory Data Analysis

After removing the null values we have merged both Rossman and Store dataset. The merged dataset consists of 1017209 observations and 18 features. Now we will begin our EDA.
Just by observing the head of the dataset and understanding the features involved in it, the following assumptions could be framed:

- There's a feature called "DayOfWeek" with the values 1-7 denoting each day of the week. There would be a week off probably Sunday when the stores would be closed and we would get low overall sales

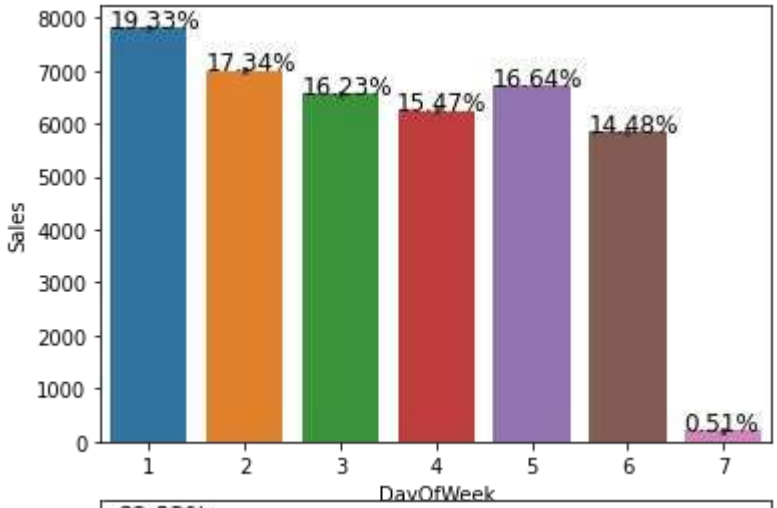
- Customers would have a positive correlation with Sales.

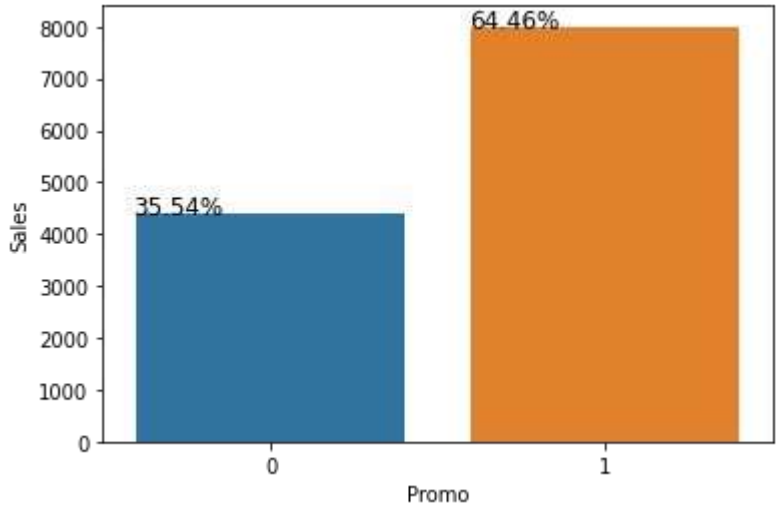
- The Store type and Assortment strategy involved would be having a certain effect on sales as well. Some premium high quality products would fetch more revenue.

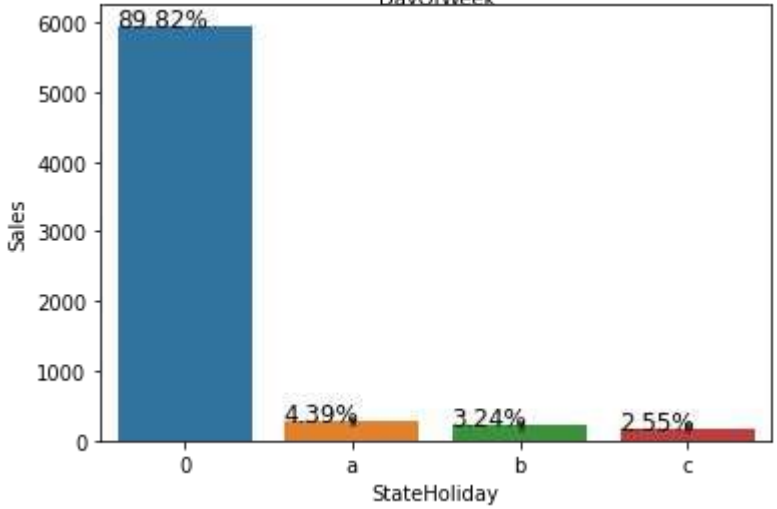
- Promotion should be having a positive correlation with Sales.

- Some stores are closed due to refurbishment, those would generate 0 revenue for that time period.

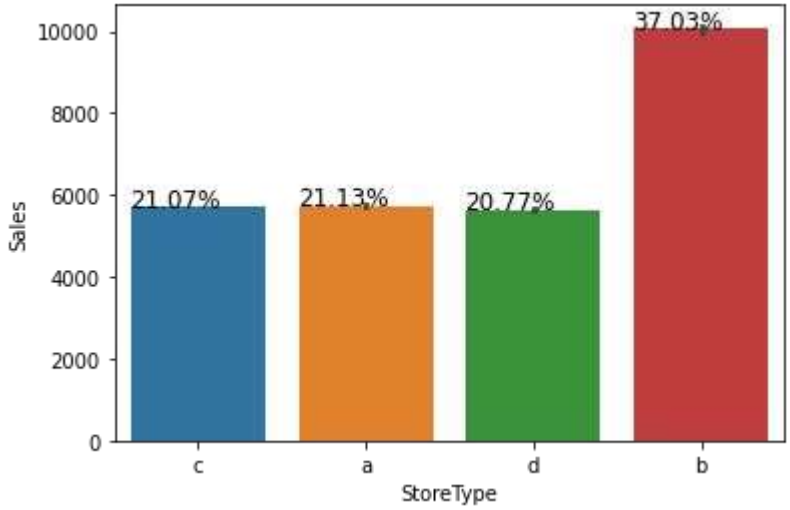
- There would be some seasonality involved in the sales pattern, probably before holidays sales would be high.

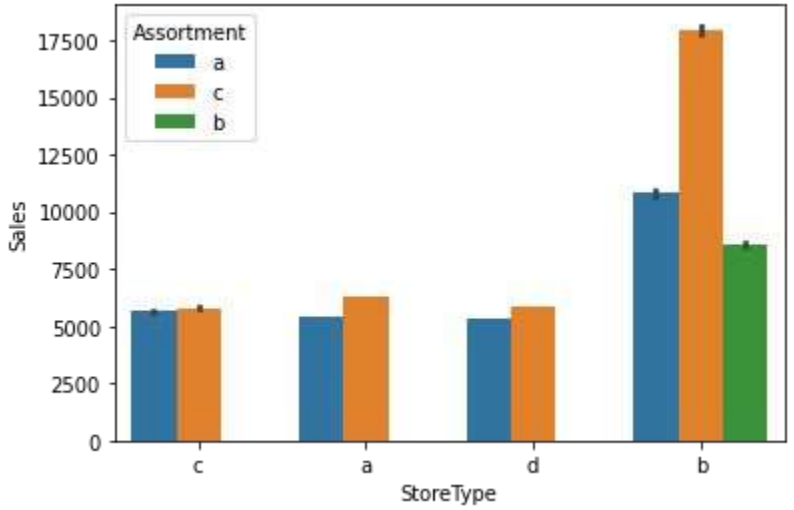






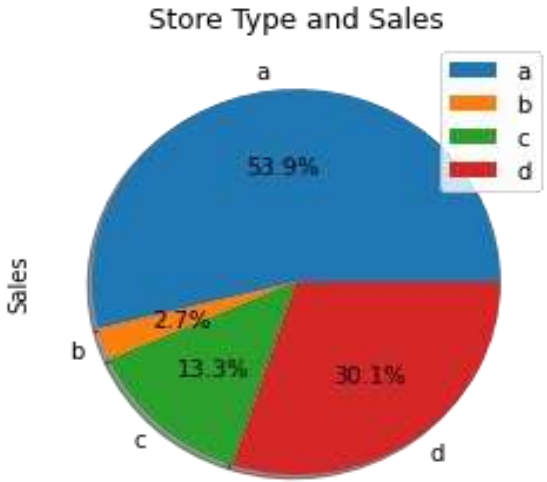
- There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week.
- Promo leads to more sales.
- Normally all stores, with few exceptions, are closed on state holidays. Lowest of Sales were seen on state holidays especially on Christmas.

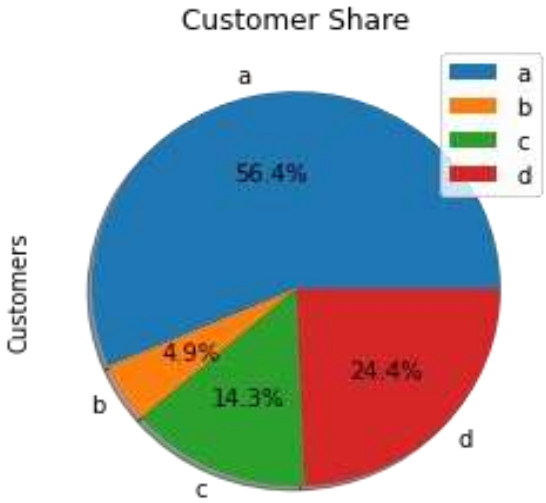


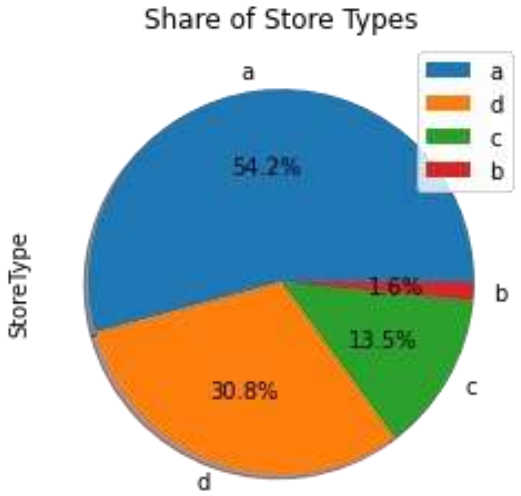


- A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. Here, it can be seen that on an average Store type B had the highest average sales.

- 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.

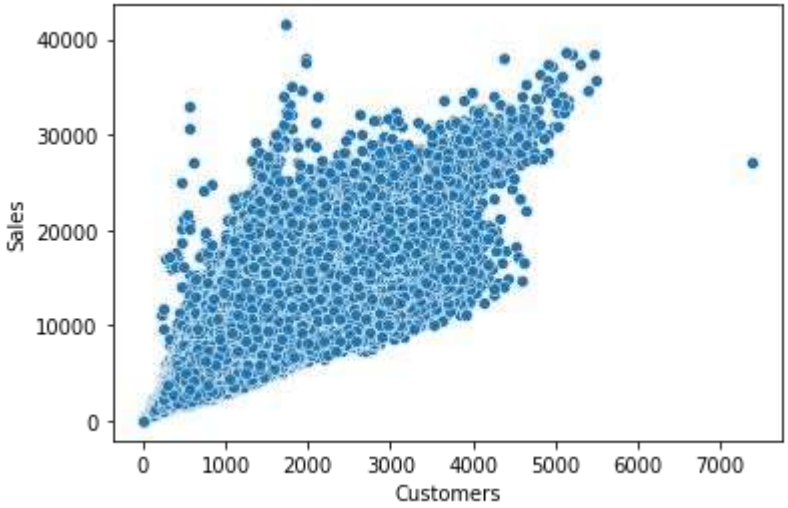


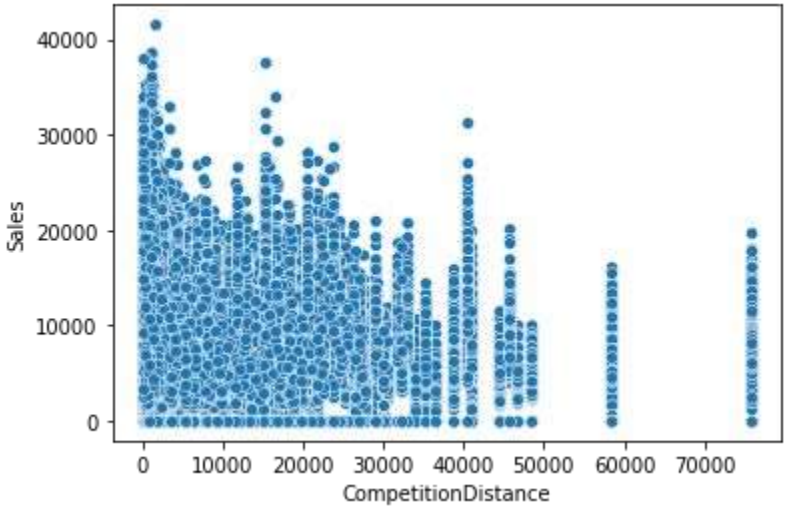


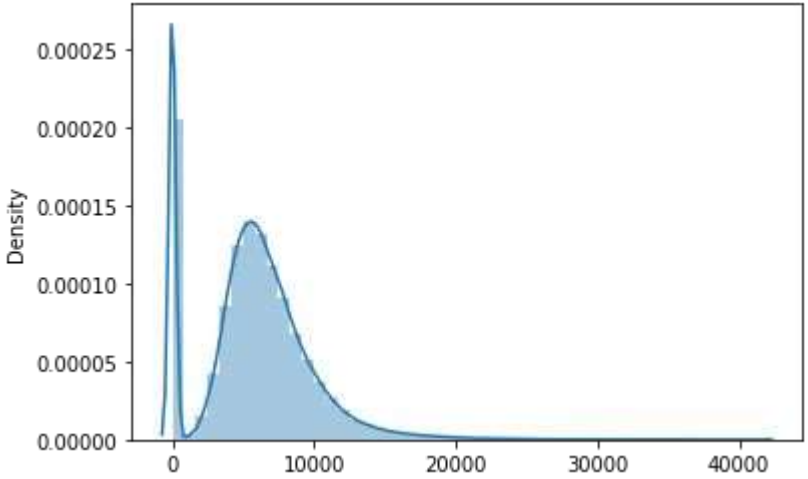


- Given the large proportion of type A stores in our dataset, further investigation made it abundantly evident that the highest sales belonged to that store type. Sales and customer share for stores of types A and C were comparable.

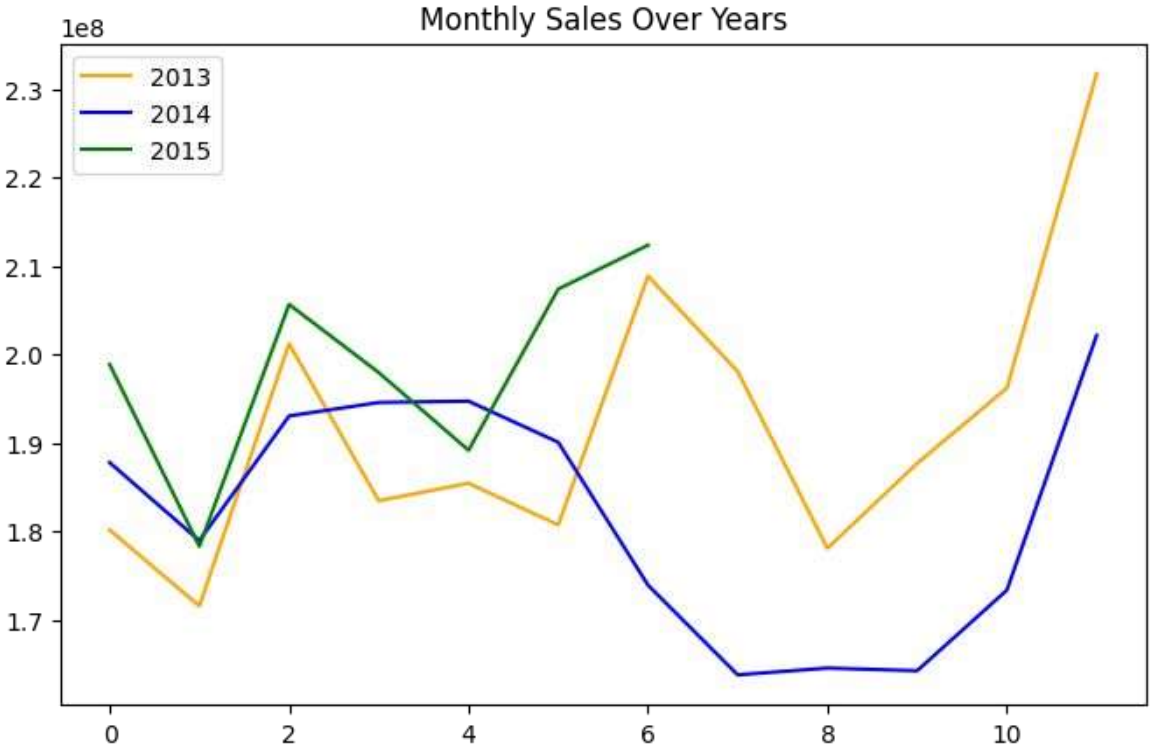
- According to the aforementioned statistics, store types "b" and "d" appear to offer a good deal of opportunity since they had higher consumer density and higher sales per customer, respectively. Because they made up the bulk of the stores, store types a and c had the highest overall income figures even if their "per customer and per store" sales figures are relatively similar. However, even though they were limited in number, stores of type B had higher average sales than the competition.







- 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 than stores far away.
- The drop in sales indicates the 0 sales accounting to the stores temporarily closed due to refurbishment.

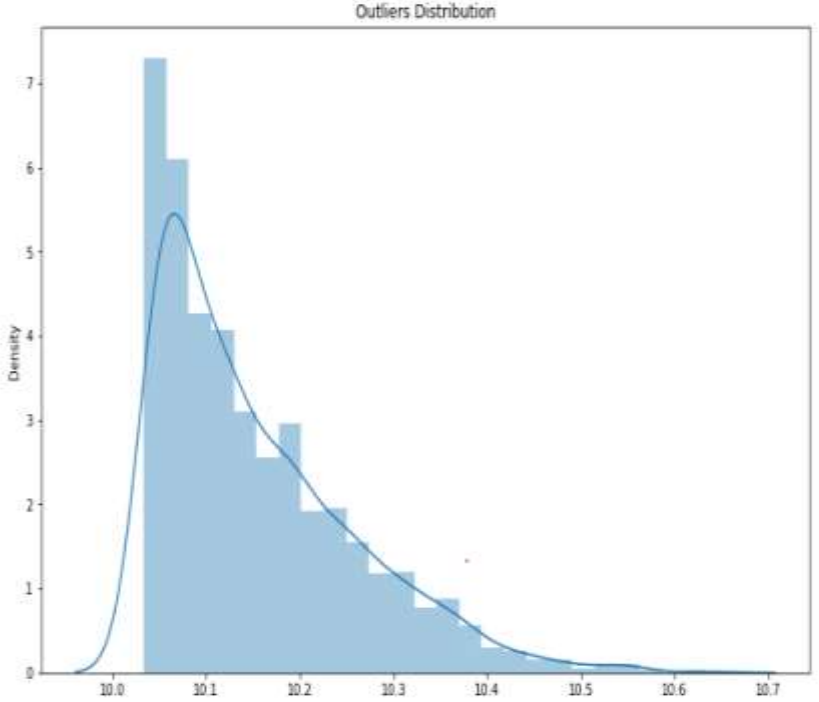


- Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September.

Outlier Detection

- An outlier is a data point in statistics that dramatically deviates from other observations. Although outliers can happen randomly in any distribution, they frequently point to measurement error or a heavy-tailed distribution in the population.

- A statistical measure called Z-score reveals how remote a data point is from the rest of the dataset. Z-score, to put it more precisely, indicates how many standard deviations an observation is from the mean. We have chosen the threshold as 3 i.e. if the z score is greater than 3 we will treat them as outliers.



As we can observe, these stores are participating in promo and hence having high sales. So these are valid outliers.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOper
1113	1114	5	2015-07-31	10.222232	3745	1	1	0	1	a	c	870.0	9.0	
5301	842	1	2015-07-27	10.235701	1493	1	1	0	0	d	c	1200.0	11.0	
20886	817	1	2015-07-13	10.210163	3437	1	1	0	0	a	a	140.0	3.0	
21183	1114	1	2015-07-13	10.245516	3592	1	1	0	0	a	c	870.0	9.0	
34563	1114	3	2015-07-01	10.206218	3788	1	1	0	0	a	c	870.0	9.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
979001	817	1	2013-02-04	10.362462	4067	1	1	0	1	a	a	140.0	3.0	
993496	817	2	2013-01-22	10.210605	7388	1	1	0	0	a	a	140.0	3.0	
994611	817	1	2013-01-21	10.330942	3900	1	1	0	0	a	a	140.0	3.0	
1009106	817	2	2013-01-08	10.241744	3862	1	1	0	0	a	a	140.0	3.0	
1010221	817	1	2013-01-07	10.381676	4065	1	1	0	0	a	a	140.0	3.0	

261 rows × 18 columns

As we can see these stores are opened on Sundays as well which can be a reason for high sales.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenS
5836	262	7	2015-07-26	10.390440	4783	1	0	0	0	b	a	1180.0	5.0	
13641	262	7	2015-07-19	10.362967	4691	1	0	0	0	b	a	1180.0	5.0	
21446	262	7	2015-07-12	10.381924	4623	1	0	0	0	b	a	1180.0	5.0	
29251	262	7	2015-07-05	10.317417	4762	1	0	0	0	b	a	1180.0	5.0	
37056	262	7	2015-06-28	10.268721	4450	1	0	0	0	b	a	1180.0	5.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
932731	262	7	2013-03-17	10.247822	4204	1	0	0	0	b	a	1180.0	5.0	
940536	262	7	2013-03-10	10.207068	4130	1	0	0	0	b	a	1180.0	5.0	
948341	262	7	2013-03-03	10.280210	4314	1	0	0	0	b	a	1180.0	5.0	
971756	262	7	2013-02-10	10.209280	4133	1	0	0	0	b	a	1180.0	5.0	
979561	262	7	2013-02-03	10.272323	4144	1	0	0	0	b	a	1180.0	5.0	

100 rows × 18 columns

Below Stores are neither running any promo nor they have their stores opened on Sunday but still having high sales because they have very large number of customers visiting the stores.

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenS
28750	876	1	2015-07-06	10.215740	1653	1	0	0	0	a	a	21790.0	4.0	
43278	909	2	2015-06-23	10.310219	1635	1	0	0	0	a	c	1680.0	9.0	
44393	909	1	2015-06-22	10.634677	1721	1	0	0	0	a	c	1680.0	9.0	
74966	262	1	2015-05-25	10.467636	4989	1	0	a	0	b	a	1180.0	5.0	
86696	842	5	2015-05-15	10.297791	1632	1	0	0	0	d	c	1200.0	11.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
916006	262	1	2013-04-01	10.414093	5013	1	0	b	0	b	a	1180.0	5.0	
918497	523	6	2013-03-30	10.234552	2901	1	0	0	0	c	c	50.0	11.0	
919088	1114	6	2013-03-30	10.245835	3944	1	0	0	0	a	c	870.0	9.0	
965364	560	6	2013-02-16	10.356091	3096	1	0	0	0	c	c	1910.0	7.0	
966479	560	5	2013-02-15	10.313708	2651	1	0	0	0	c	c	1910.0	7.0	

68 rows × 18 columns

From the above analysis we can conclude that there is no absurd behaviour and hence we are not treating any outliers.

Feature Engineering and Feature Selection

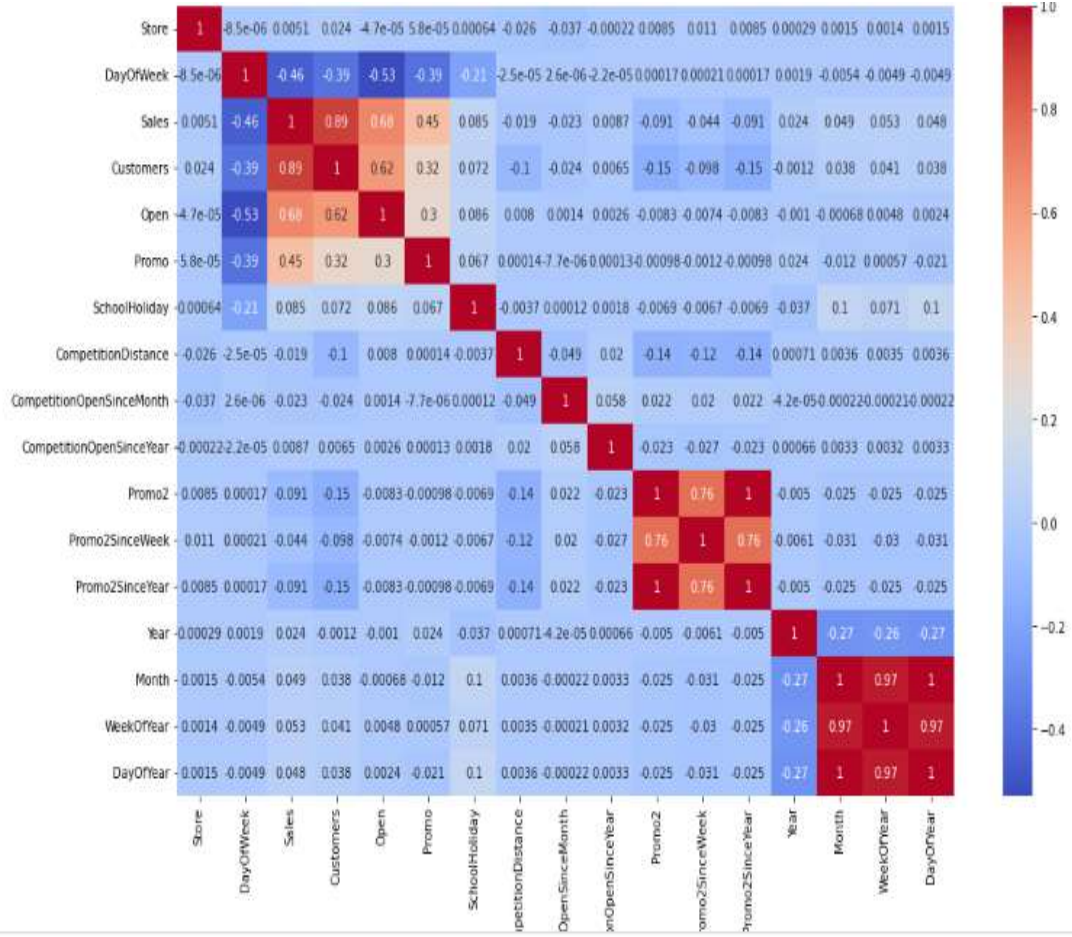
Process followed in feature engineering and feature selection are summarised below –

1) First step was to check the distribution of our Sales (Y variable). So for this we dropped all the datasets where stores were closed (sales will always be 0) and then plotted the distribution. We found out that Sales is positively skewed and hence we applied log transform to convert it to normal distribution.

2) We defined a new feature ‘CompetitionOpen’ by subtracting CompetitionOpenSinceYear and CompetitionOpenSinceMonth from our Date column.

3) Then we found the VIF and plotted the heatmap to see the correlation among the features. We dropped off Year, month, WeekOfYear, DayOfYear, CompetitionOpenSinceYear, CompetitionOpenSinceMonth, promo2, promo2SinceWeek and promo2SinceYear to remove the multicollinearity.

Heatmap and VIF



	variables	VIF
0	DayOfWeek	3.1338
1	Customers	5.0742
2	Open	7.9294
3	Promo	1.8839
4	SchoolHoliday	1.2403
5	CompetitionDistance	1.5244
6	Promo2	2.0196
7	WeekOfYear	3.3734
8	CompetitionOpen	1.4610

Label Encoding

1) We have done the hot labe encoding for StoreType, Assortment and DayOfWeek.
2) We have replaced the a, b and c StateHolidays with 1.

Train Test Split

1) We have split the dataset on the basis of the Date column.
2) Train dataset consists of datapoints which lie between ‘2013-01-01’ and ‘2015-06-14’.
3) Test dataset consists of datapoints which lie between ‘2015-06-15’ and ‘2015-07-31’.

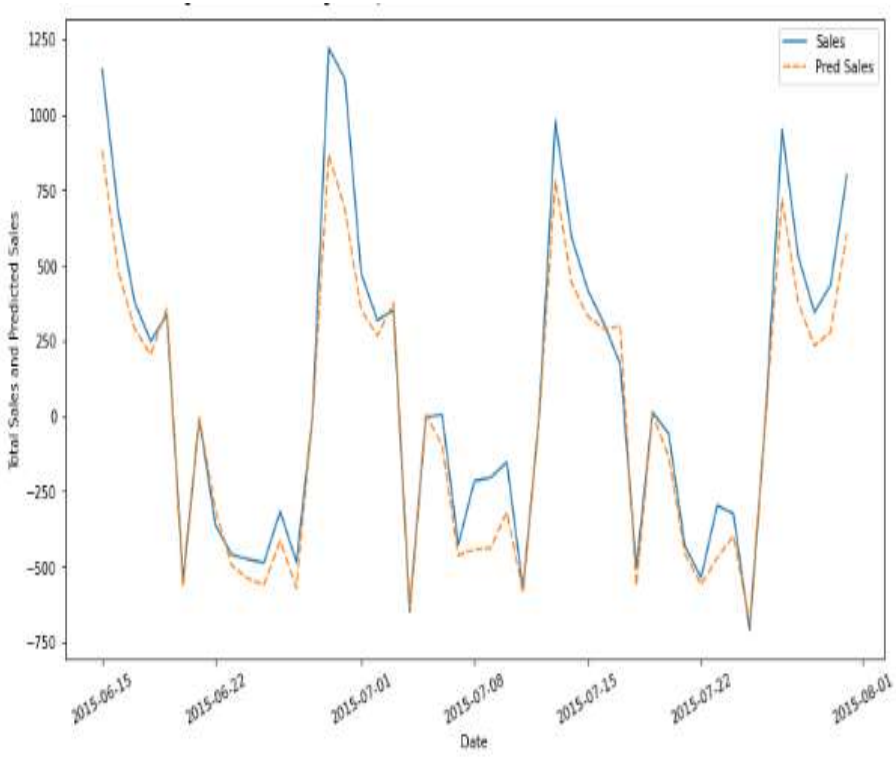
Scaling

1) We have applied the Standard Scaler on test and train dataset.

Linear Regression

1) Due to the fact that our dataset primarily consists of categorical columns with a few continuous characteristics, such as customers and competition distance, linear regression is not particularly effective at forecasting the final result.

2) Overall, if we look at the graph of anticipated sales vs. real sales, the most of the predicted points are relatively close to the actual values of sales, but there are a small number of spots where it is not able to forecast accurately. Applying the Decision Tree approach could help solve this issue.

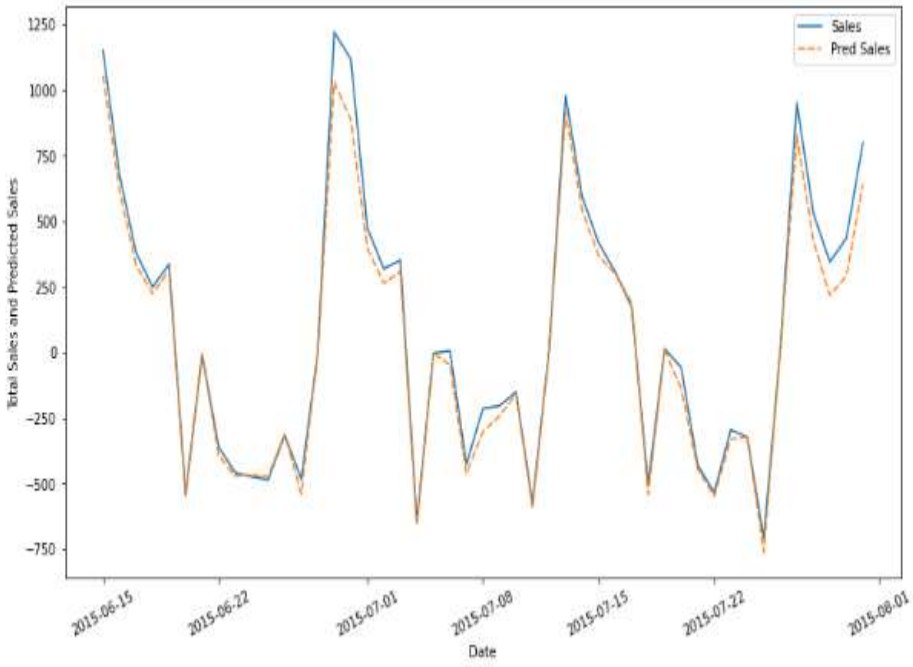


	MAE_train	MSE_train	RMSE_train	R2_train	Adj_r2_train	train_score	MAE_test	MSE_test	RMSE_test	R2_test	Adj_r2_test	test_score
Linear Regression	0.377706	0.253608	0.503595	0.746392	0.746385	0.746392	0.379612	0.246445	0.496433	0.738373	0.738247	0.738373

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 effect which can handle non-linear data sets effectively.

- The results show that a simple decision tree is performing pretty well on the validation set but it has completely overfitted the train set. It's better to have a much more generalized model for future data points.



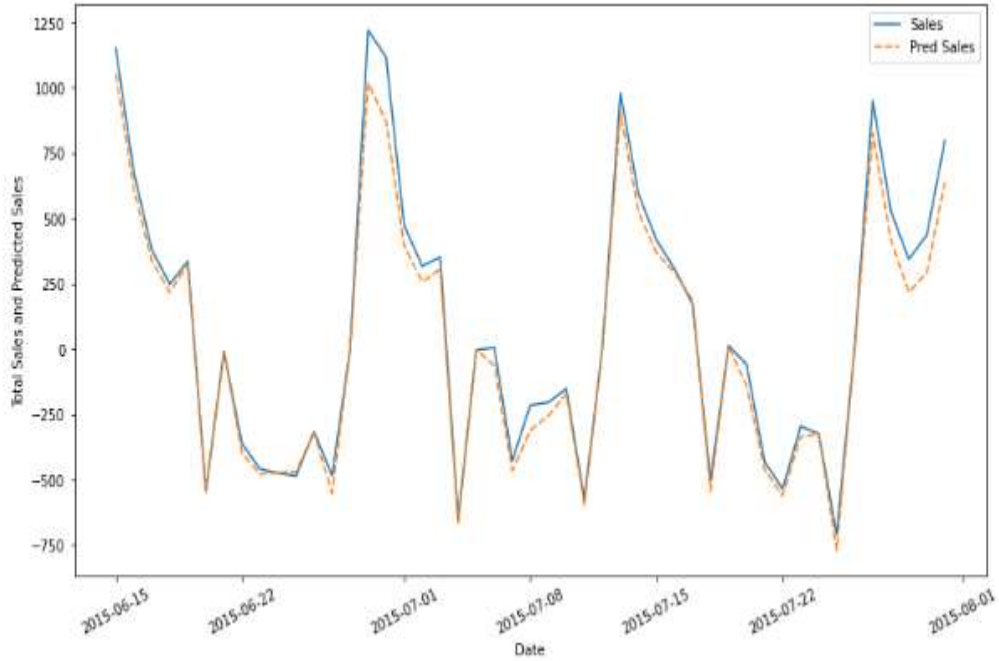
	MAE_train	MSE_train	RMSE_train	R2_train	Adj_r2_train	train_score	MAE_test	MSE_test	RMSE_test	R2_test	Adj_r2_test	test_score
DecisionTreeRegressor	0.02458	0.003385	0.05818	0.996615	0.996615	0.996615	0.187829	0.066215	0.257322	0.929706	0.929673	0.929706

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.

- To prevent overfitting, we built random forest model. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

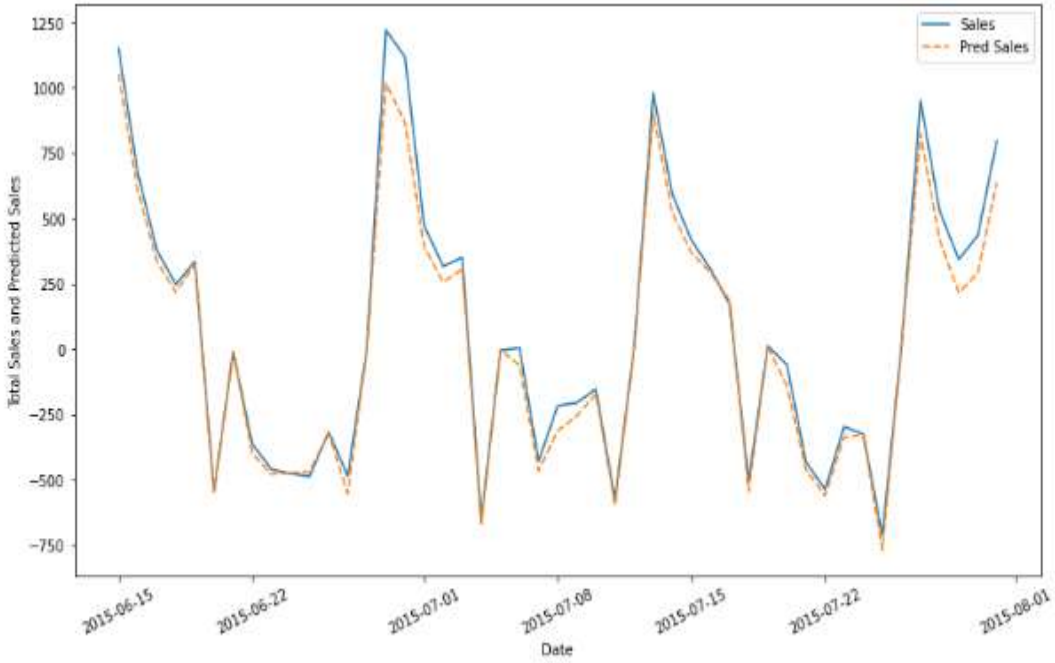
- Random Forest Regressor results were much better than our baseline model with a test R^2 of 0.9623.



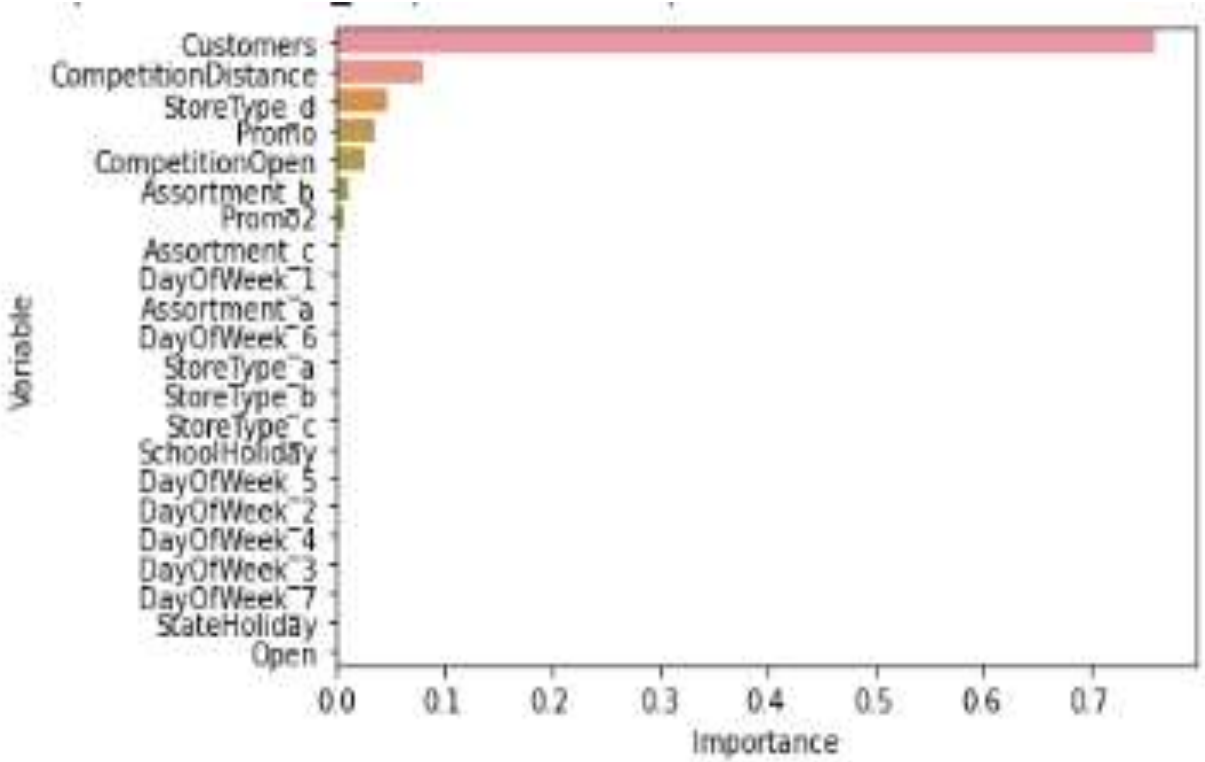
Random Forest Regressor	0.045540	0.003920	0.062607	0.996080	0.996080	0.996080	0.142370	0.035497	0.188406	0.962317	0.962299	0.962317

Random Forest Hyperparameter Tuning

- The maximum R^2 was seen in tuned Random Forest model with the value 0.962559.
- This 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.



Random Forest 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.

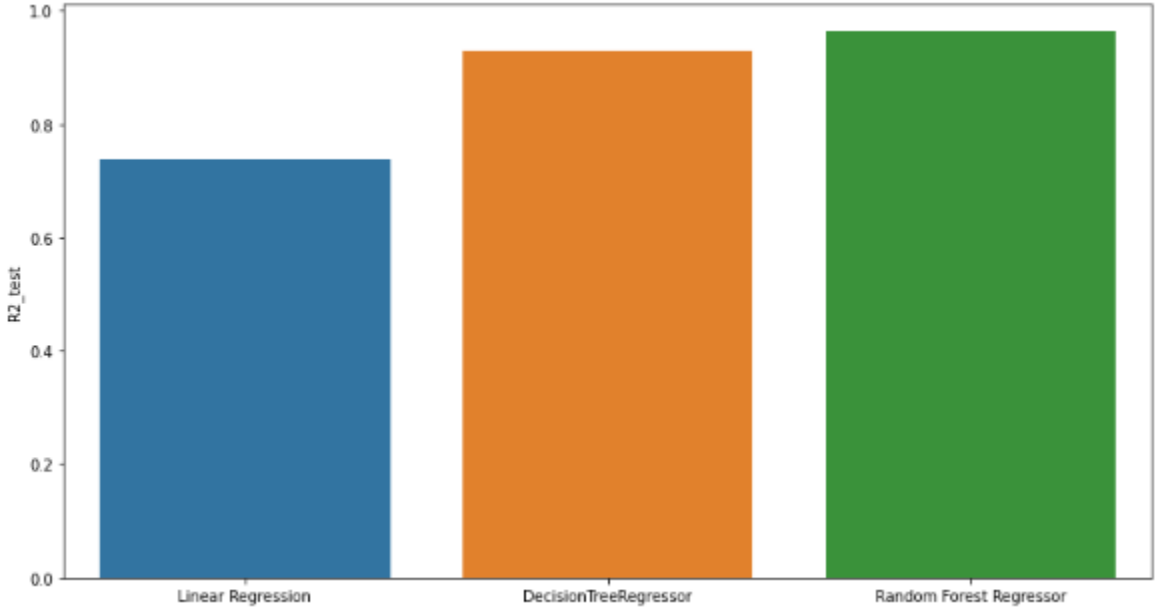
- The baseline model completely overfitted the data with a train R^2 of 0.996 and test R^2 of 0.9297.

- To prevent overfitting, we built random forest model. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random Forest Regressor results were much better than our baseline model with a test R^2 of 0.9623.

- This indicates that the improvement in the model performance was 3.508 % than the baseline model.

- Tuning the hyperparameters gave the best results with a test R^2 of 0.9625 which was only 0.023% improved from a simple random forest model. It signifies maxed out performance by the model on the given data.

	MAE_train	MSE_train	RMSE_train	R2_train	Adj_r2_train	train_score	MAE_test	MSE_test	RMSE_test	R2_test	Adj_r2_test	test_score
Linear Regression	0.377706	0.253608	0.503595	0.746392	0.746385	0.746392	0.379612	0.246445	0.496433	0.738373	0.738247	0.738373
DecisionTreeRegressor	0.024580	0.003385	0.058180	0.996615	0.996615	0.996615	0.187829	0.066215	0.257322	0.929706	0.929673	0.929706
Random Forest Regressor	0.045540	0.003920	0.062607	0.996080	0.996080	0.996080	0.142370	0.035497	0.188406	0.962317	0.962299	0.962317



Store wise Sales Predictions

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

Date	Sales	Pred_Sales
2015-06-15	9108.0	9104.592660
2015-06-15	11683.0	10197.404187
2015-06-15	10510.0	9723.873581
2015-06-15	15664.0	13695.238347
2015-06-15	5511.0	5443.496823
2015-06-15	11209.0	10796.210505
2015-06-15	10297.0	9850.708161
2015-06-15	13773.0	12607.314376
2015-06-15	11348.0	10099.203879
2015-06-15	5982.0	6081.862064

Conclusion and Recommendations:

Some important conclusions drawn from the analysis are as follows:

1) There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week.

2) The positive effect of promotion on Customers and Sales is observable.

3) 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.

4) 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.

5) The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.

6) Decision tree was chosen as baseline model considering our features were mostly categorical with few having continuous importance.

7) Random Forest shows improvement of 3.508% as compared to Decision tree.

8) Random Forest Tuned Model gave the best results and only 0.023% improvement was seen from the basic random forest model which indicates that all the trends and patterns that could be captured by these models without overfitting were done and maximum level of performance achievable by the model was achieved.

Recommendations:

1) More stores should be encouraged for promotion.

2) Store type B should be increased in number.

3) There's a seasonality involved, hence the stores should be encouraged to promote and take advantage of the holidays.

References:

- Machine Learning Mastery
- GeeksforGeeks
- Towards Data Science Blogs
- Built in Data Science Blogs
- Scikit-Learn Org
- StackOverFlow
- StackExchange
- Medium

Thank you