Capstone Project

(Supervised ML- Regression)

Retail Sales Prediction

Project by:

Nethinti Ramakrishna

ramakrishna09nethinti@gmail.com

Content

- Problem Statement
- Retail Sales Prediction
- Data Summary
- Approach
- Exploratory Data Analysis
- Outlier Detection
- Modeling:

  - Baseline Model - Decision Tree
  - Random Forest
  - Random forest Hypertuning Parameters
  - Feature Importance
- 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 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

- **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**
  - Data Cleaning
  - Missing Data Handling
  - Merging the Datasets
- **Exploratory Data Analysis**
  - Hypotheses
  - Categorical Features
  - Continuous Features
  - EDA Conclusion and Validating Hypotheses
- **Data Manipulation**
  - Feature Engineering
  - Outlier Detection and Treatment
  - Feature Scaling
  - Categorical Data Encoding
- **Modeling**
  - Train Test Split
  - Baseline Model - Decision Tree
  - Random Forest Model
  - Random Forest Hyperparameter Tuning
  - Random Forest Feature Importance
- **Model Performance and Evaluation**
  - Visualizing Model Performances
  - Random Forest vs Baseline Model
  - Random Forest Tuned vs Baseline and Random Forest Models
- **Store wise Sales Predictions**
- **Conclusion and Recommendations**

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.

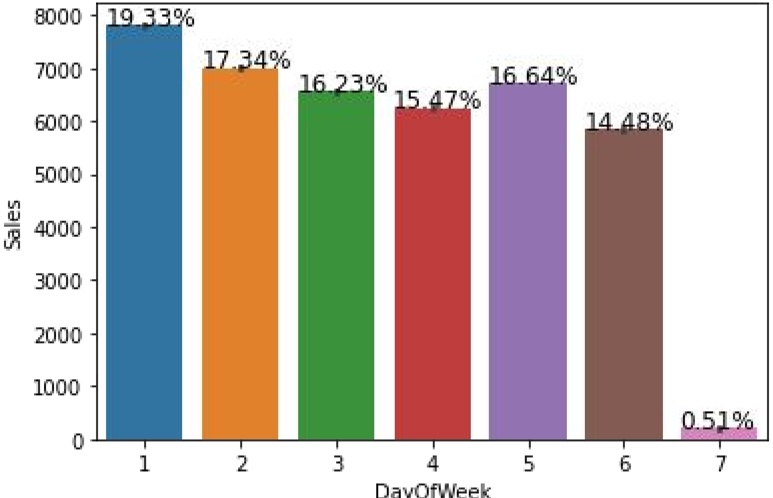
- Customers would have a positive correlation with Sales.

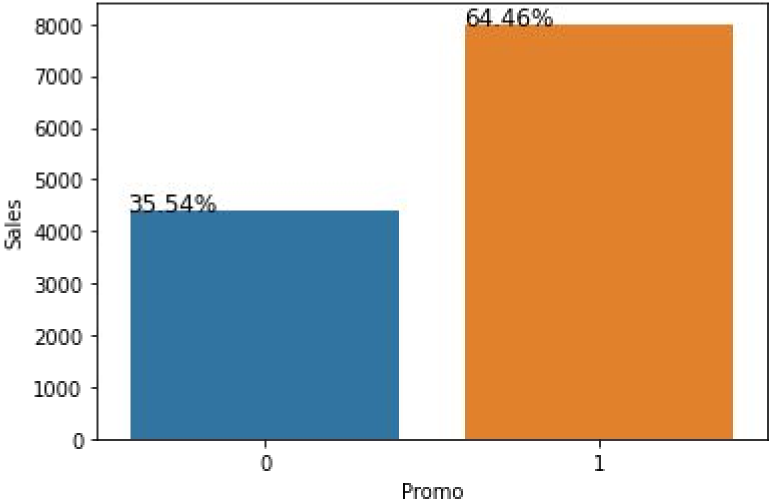
- The Store type and Assortment strategy involved would be having a certain effect on sales as well. Some premium high quality products would fetch more revenue.

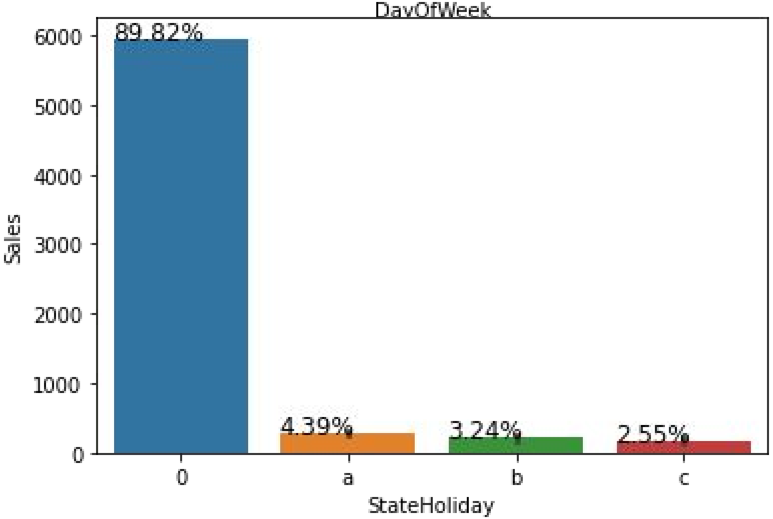
- Promotion should be having a positive correlation with Sales.

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

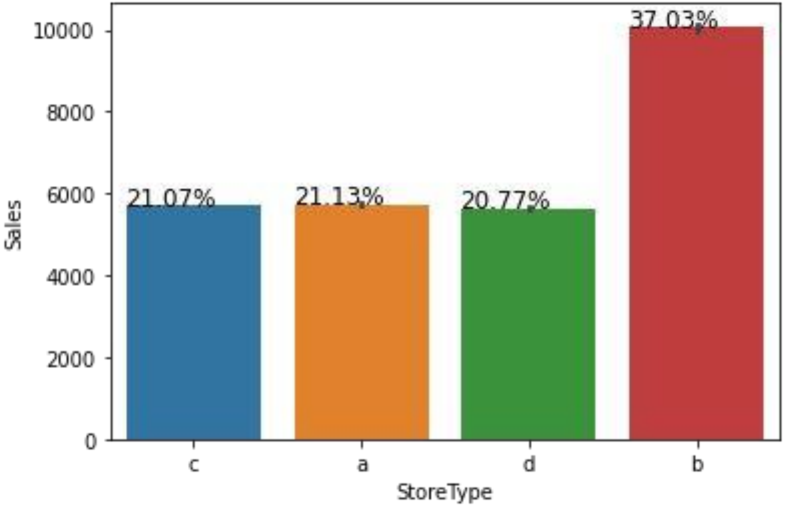
- There would be some seasonality involved in the sales pattern, probably before holidays sales would be high.

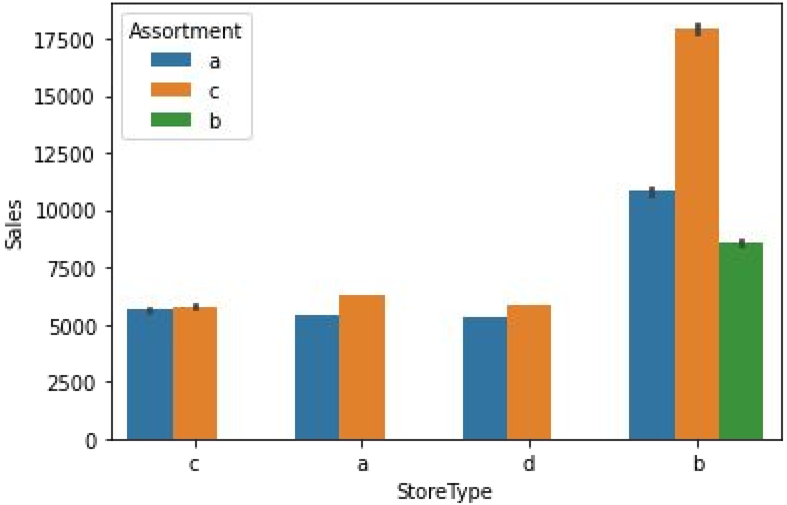






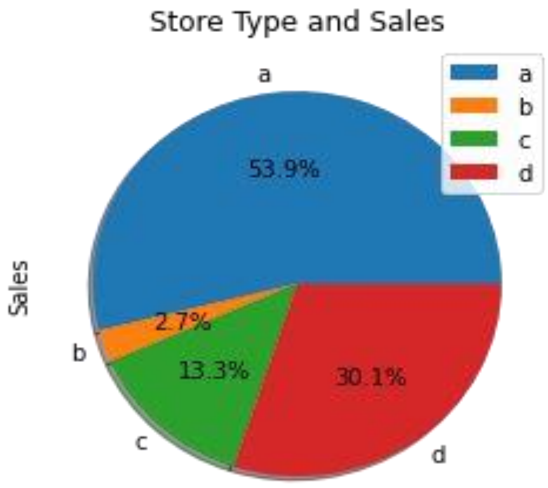
- There were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week.
- Promo leads to more sales.
- Normally all stores, with few exceptions, are closed on state holidays. Lowest of Sales were seen on state holidays especially on Christmas.



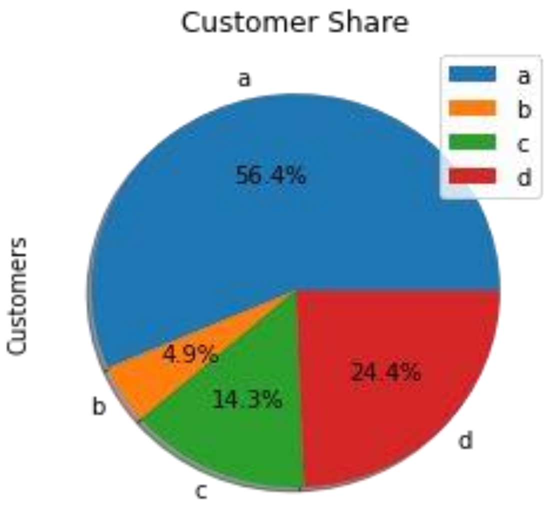


- A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. Here, it can be seen that on an average Store type B had the highest sales. There has to be something different about this store type.

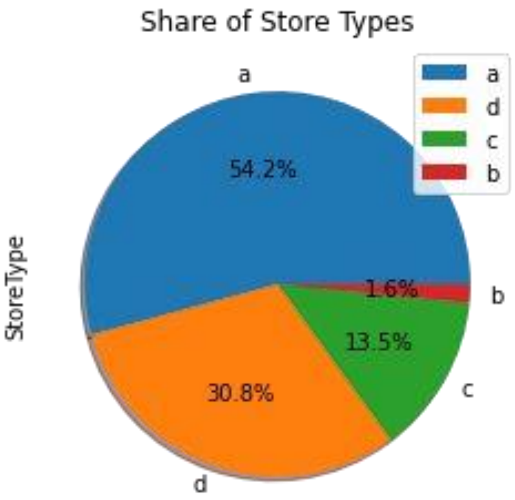
- Next it can be seen that the store types a, c and d have only assortment level a and c. On the other hand the store type b has all the three kinds of assortment strategies.



Store Type and Sales



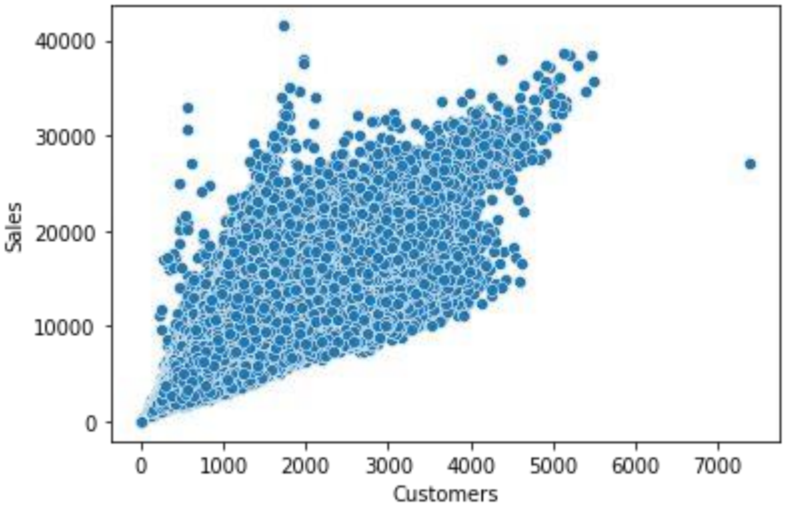
Customer Share

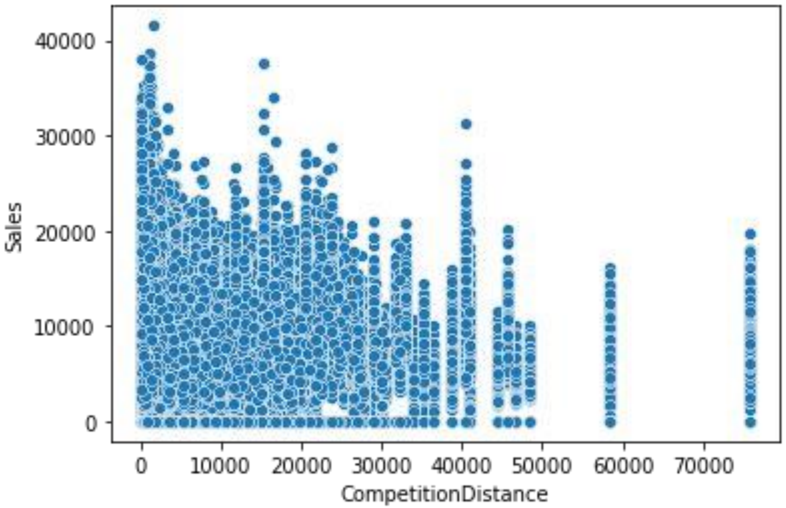


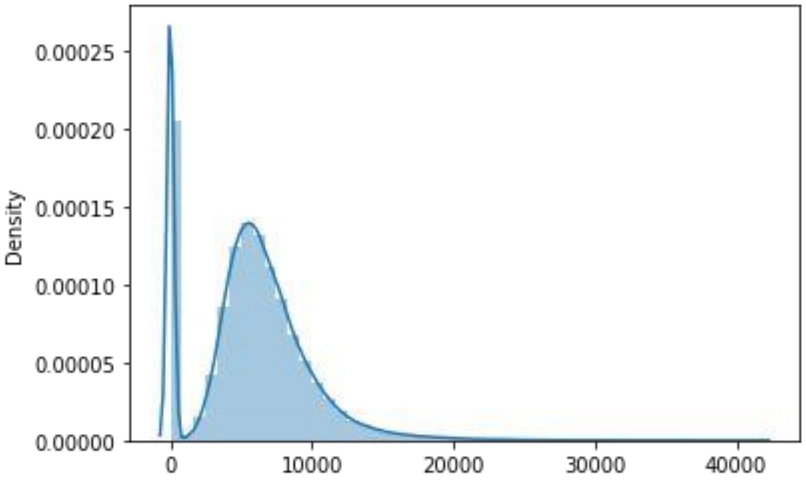
Share of Store Types

- 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 '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.



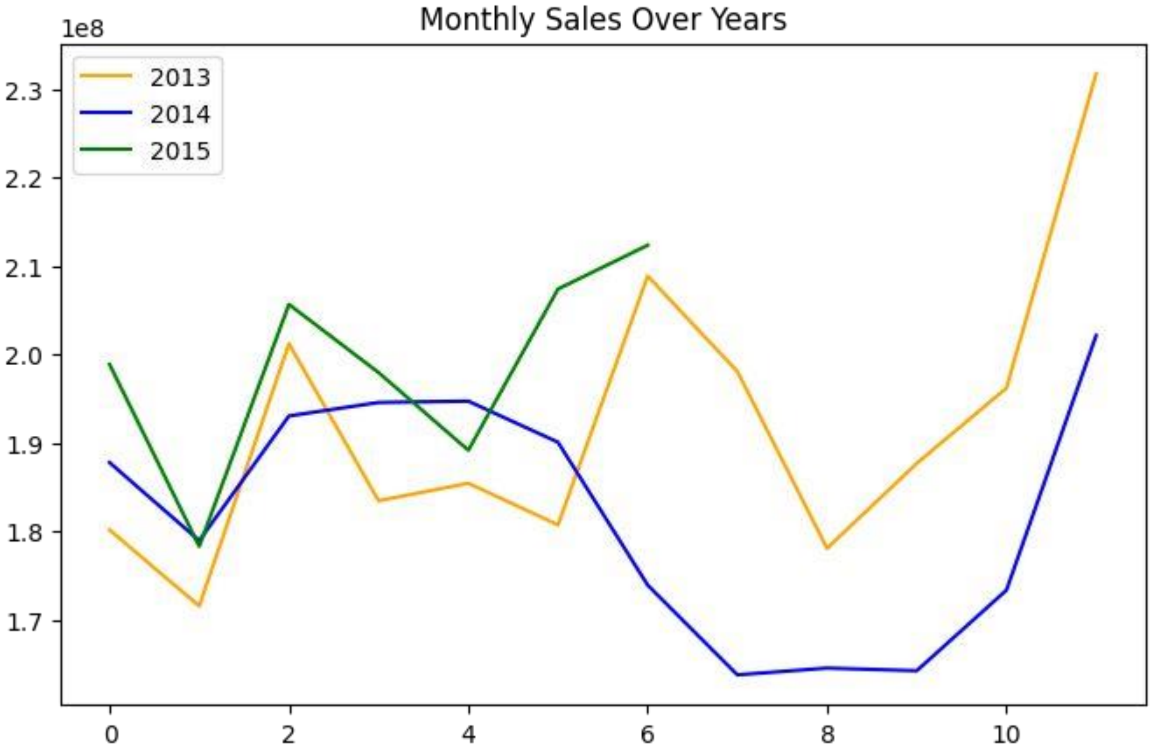




- It's pretty obvious that there is going to be a positive correlation between customers and sales. There are a few outliers.

- 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.

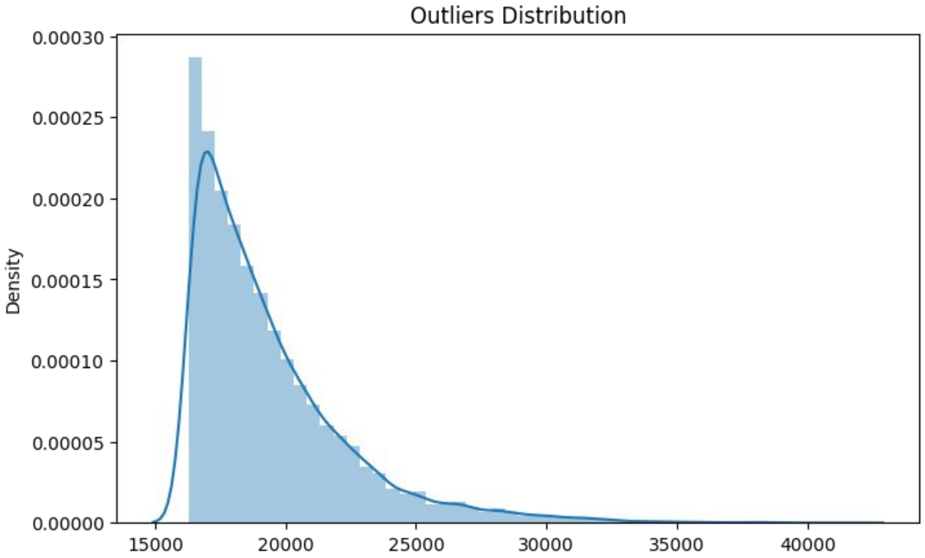


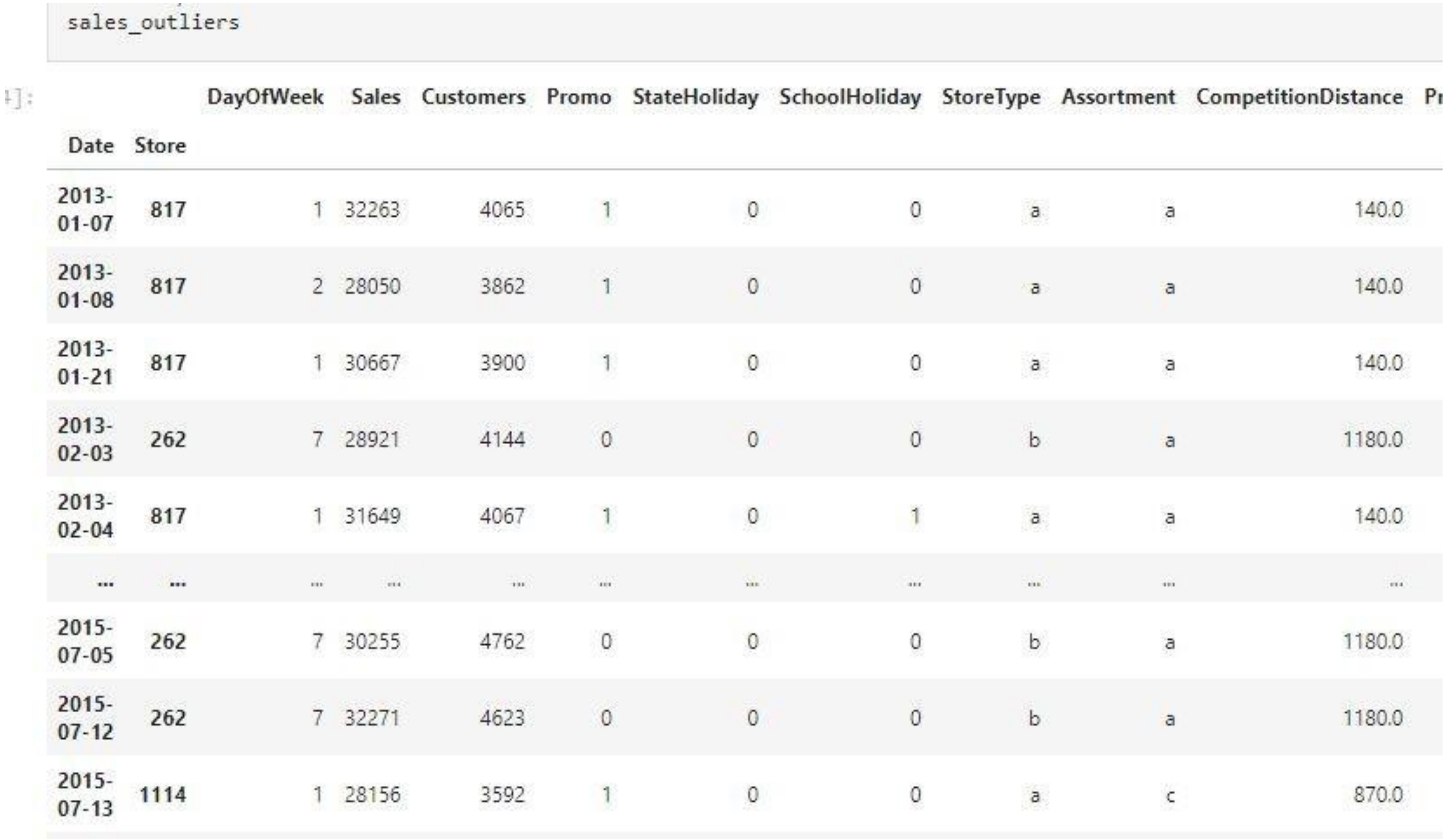
- Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.

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.

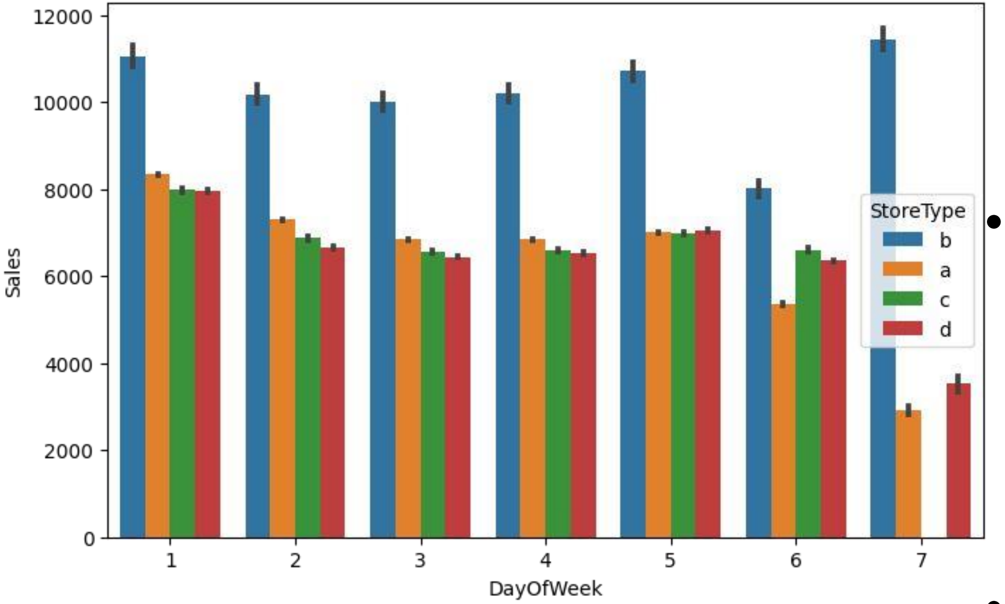




```
sales_outliers
```

4]:

		DayOfWeek	Sales	Customers	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	Pr
Date	**Store**										
2013-01-07	**817**	1	32263	4065	1	0	0	a	a	140.0	
2013-01-08	**817**	2	28050	3862	1	0	0	a	a	140.0	
2013-01-21	**817**	1	30667	3900	1	0	0	a	a	140.0	
2013-02-03	**262**	7	28921	4144	0	0	0	b	a	1180.0	
2013-02-04	**817**	1	31649	4067	1	0	1	a	a	140.0	
...	**...**	...	...	...	...	...	...	...	...	...	
2015-07-05	**262**	7	30255	4762	0	0	0	b	a	1180.0	
2015-07-12	**262**	7	32271	4623	0	0	0	b	a	1180.0	
2015-07-13	**1114**	1	28156	3592	1	0	0	a	c	870.0	



- 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 kinds of assortments available is probably the reason why it had higher average sales than any other store type.

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:

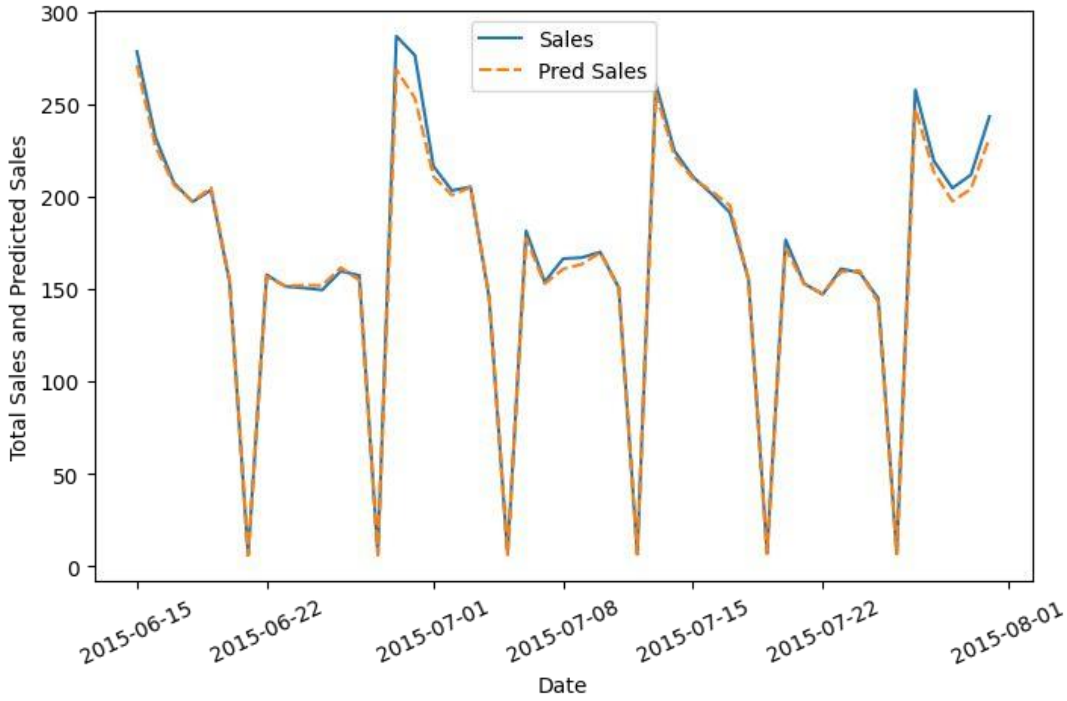
- A multivariate time series relation with sales and hence a linear relationship cannot be assumed in this analysis. This kind of dataset has patterns such as peak days, festive seasons etc which would most likely be considered as outliers in simple linear regression.

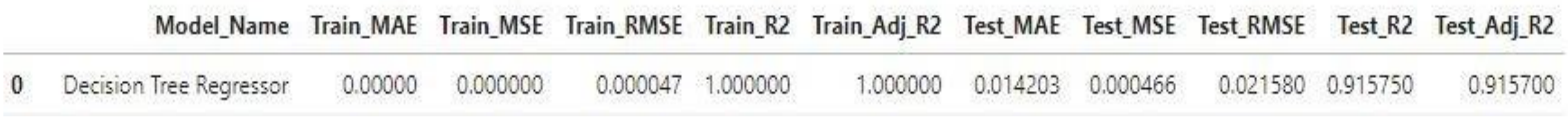
- Having X columns with 30% continuous and 70% categorical features. Business prefers the model to be interpretable in nature and decision based algorithms work better with categorical data.

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.





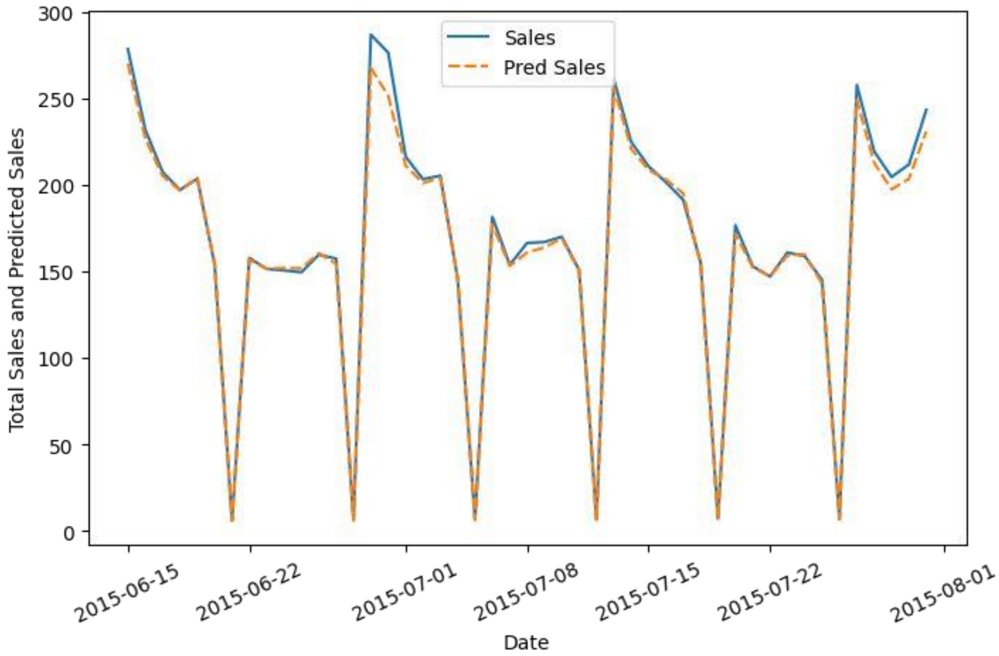
	Model_Name	Train_MAE	Train_MSE	Train_RMSE	Train_R2	Train_Adj_R2	Test_MAE	Test_MSE	Test_RMSE	Test_R2	Test_Adj_R2
0	Decision Tree Regressor	0.00000	0.000000	0.000047	1.000000	1.000000	0.014203	0.000466	0.021580	0.915750	0.915700

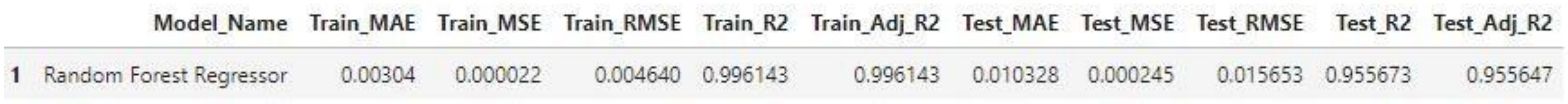
Random Forest

- 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 0.955673.



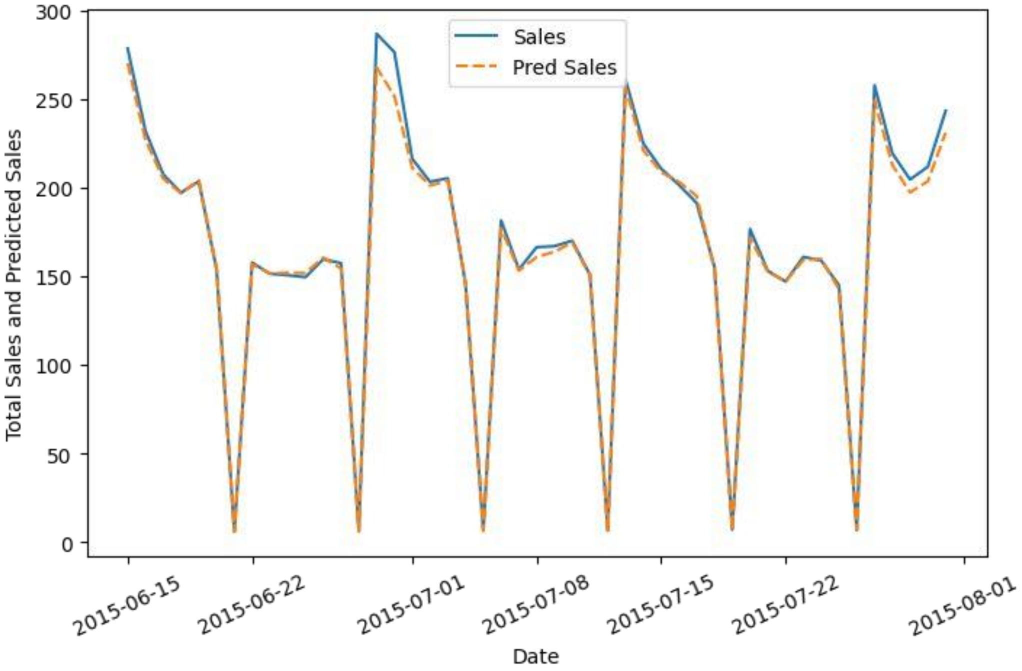


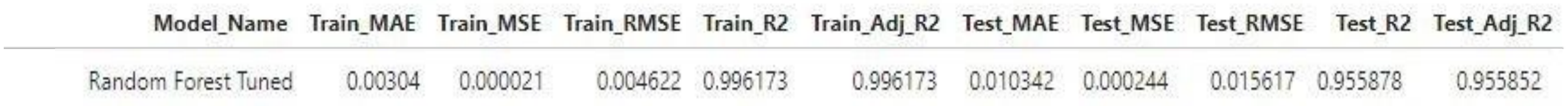
	Model_Name	Train_MAE	Train_MSE	Train_RMSE	Train_R2	Train_Adj_R2	Test_MAE	Test_MSE	Test_RMSE	Test_R2	Test_Adj_R2
1	Random Forest Regressor	0.00304	0.000022	0.004640	0.996143	0.996143	0.010328	0.000245	0.015653	0.955673	0.955647

Random Forest Hyperparameter Tuning

- The maximum R^2 was seen in tuned Random Forest model with the value 0.955878 which was only 0.021% improved from a simple random forest model.

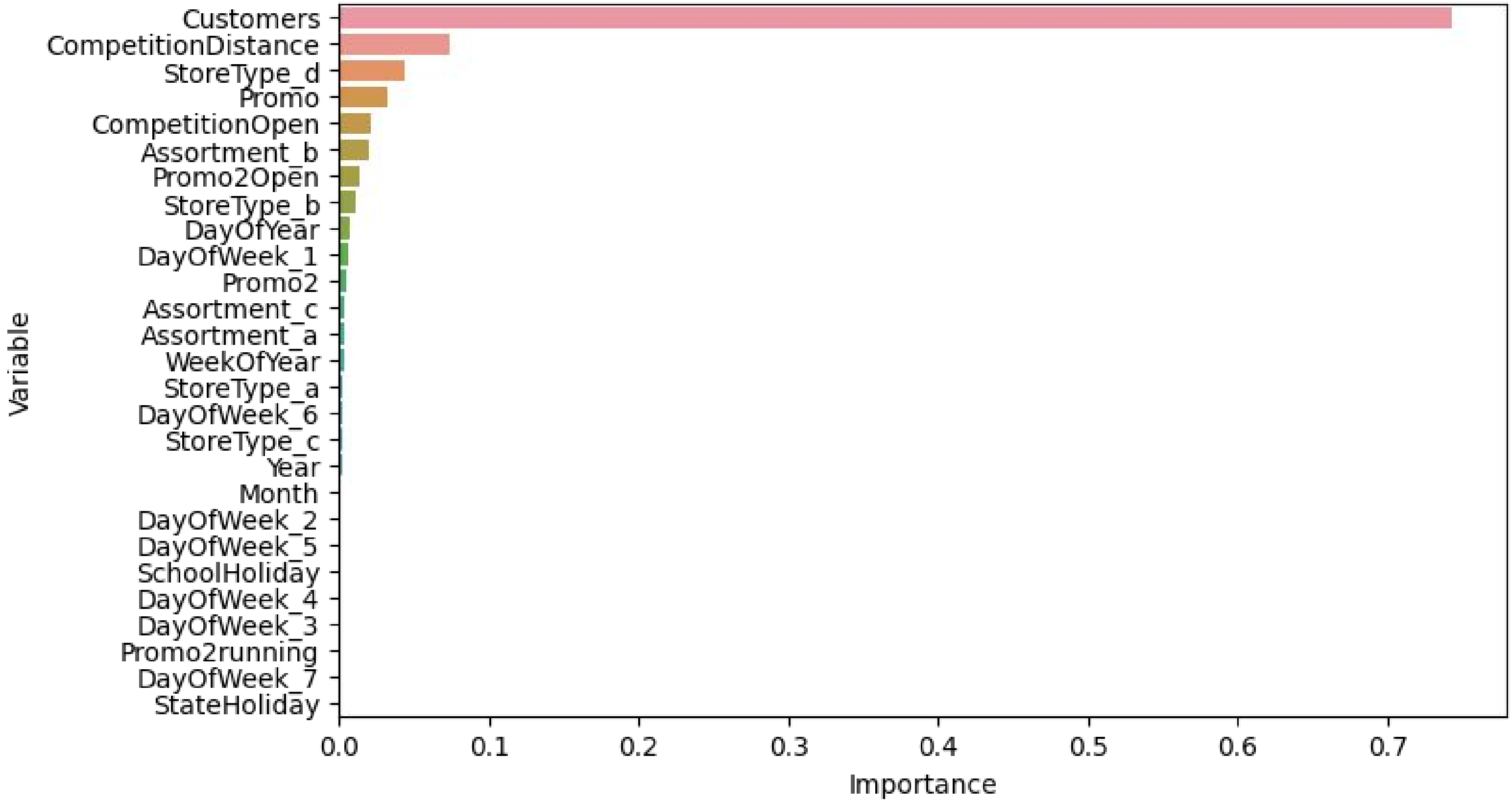
- This indicates that all the trends and patterns that could be captured by these models without overfitting were done and maximum level of performance achievable by the model was achieved.





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 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 1 and test R^2 of 0.91575.

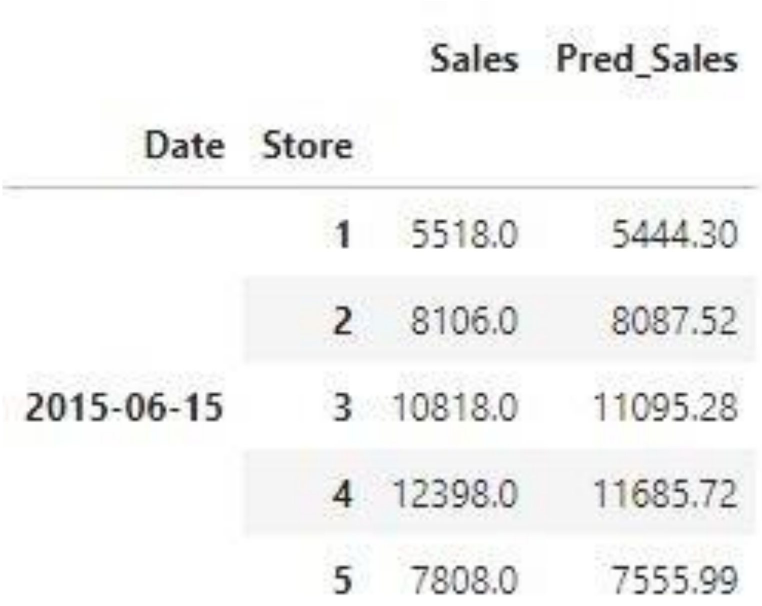
- To prevent overfitting, we built random forest model. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random Forest Regressor results were much better than our baseline model with a test R^2 of 0.955673.

- This indicates that the improvement in the model performance was 4.36% than the baseline model.

- Tuning the hyperparameters gave the best results with a test R^2 of 0.955878 which was only 0.021% improved from a simple random forest model. It signifies maxed out performance by the model on the given data.

Store wise Sales Predictions

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



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

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 kms and had more sales than stores far away probably indicating competition in busy locations vs remote locations.
- Store type B though being few in number had the highest sales average. The reasons include all three kinds of assortments specially assortment level b which is only available at type b stores and being open on sundays as well.
- The outliers in the dataset showed justifiable behaviour. The outliers were either of store type b or had promotion going on which increased sales.

Recommendations:

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

References:

- AlmaBetter
- GeeksforGeeks
- DataCamp
- Coursera
- Kaggle
- freecodecamp