Capstone Project
Retail Sales prediction

Nitesh Bhowmick

Content

Problem Statement

- Prediction based on sales
- Prediction based on sales on dependent variable
- Prediction based on store type
- Prediction based on sales between assortment and store type
- Prediction based on state holiday and school holiday
- Prediction based on day of week and open promo

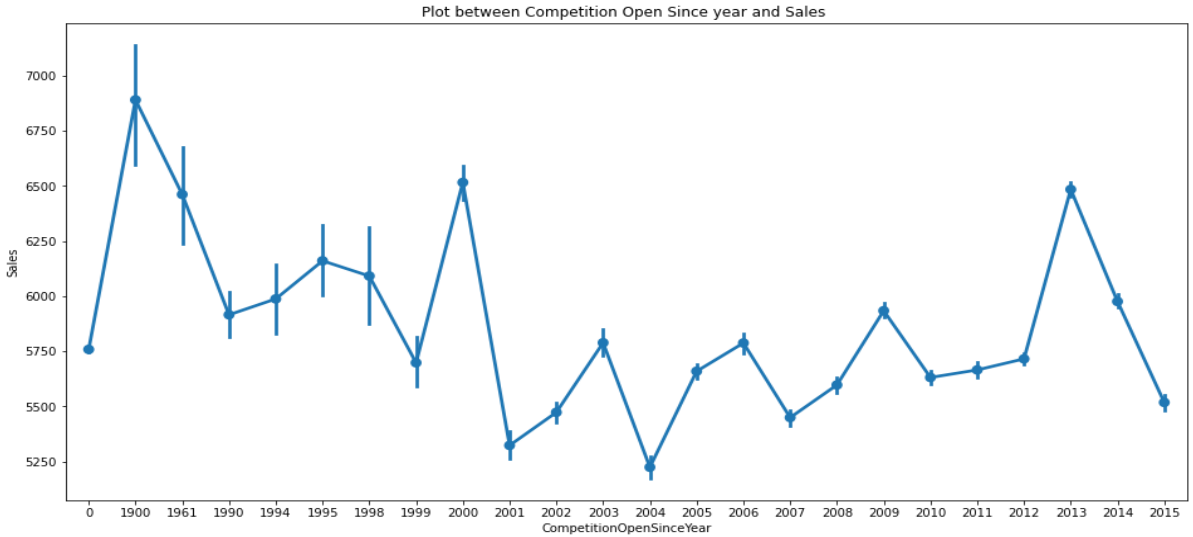
Data Summary

- In the Rossmann sales prediction project there is a dataset which contains sales information
- The sales column contain 172817 rows with 0 sale. So we created a new data frame in which we removed 0 sales rows and tried train our model we used various algorithms and got accuracy score around 74%
- The total dataset sale =0 rows . So we trained and another model using various algorithms accuracy near about 92%which is far better than previous model.
- The removing sales =0 rows actually removes lot of information from dataset as it has 172817 rows which is quite large

Exploratory Data Prediction

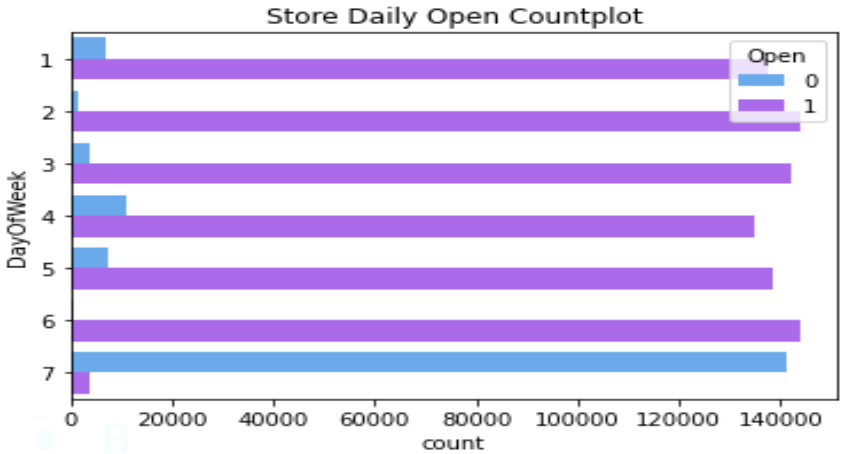
- Exploratory Data Prediction is also known as EDA, is the process of interpreting datasets by summarizing their key properties and frequently them
- EDA refers to the critical process of performing initial investigation on datasets so as discover the patterns, to spots anomalies , to hypothesis, and to check assumptions with the help of summary statics and graphical representation
- In EDA, plotting option include box plot, line plots , scatter plots and many more.

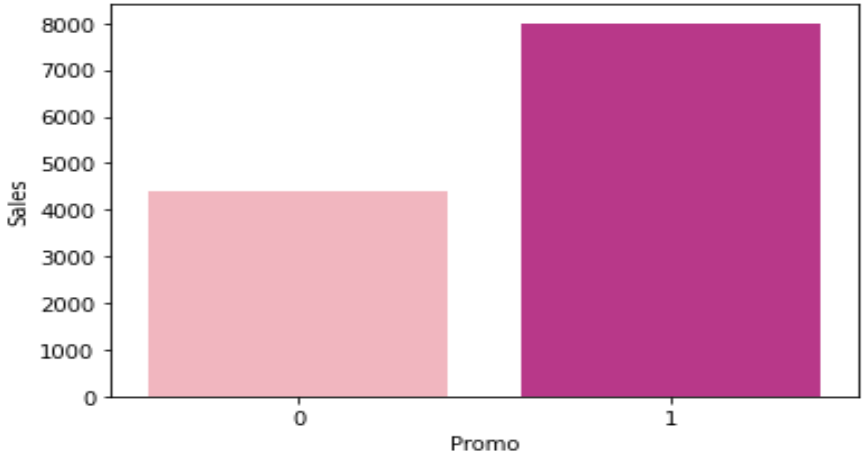
Sales



From the plot we can see sales are high during the year 1900. as there are very few store were operated of Rossmann so there is less competition and sales are high but the pass year no of store increased

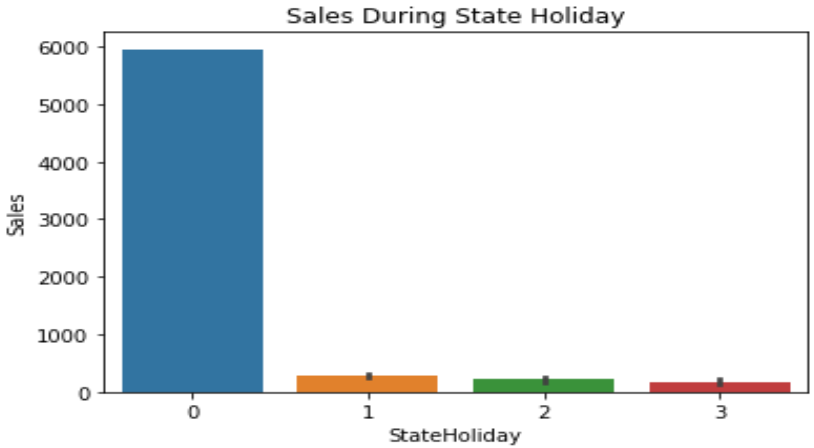
Day of week & open promo

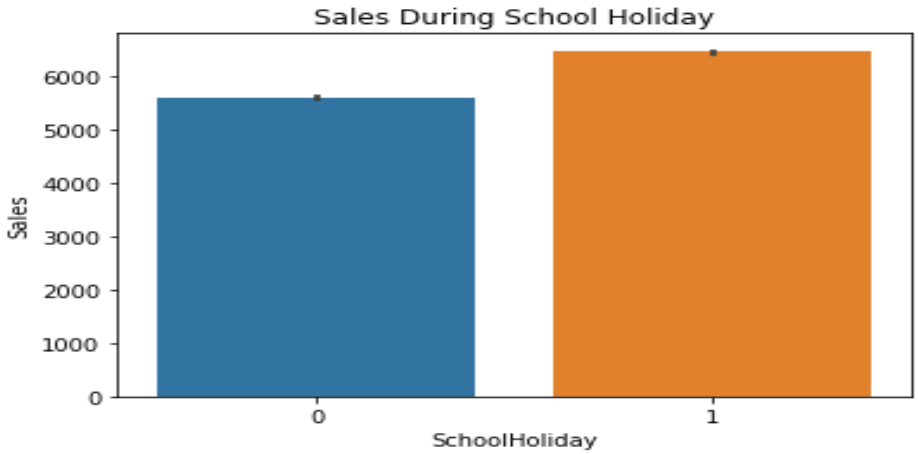




- Bar plot b/w promo and sales shows the effect of promotion on sales here 0 represents the store which did not opt for promo and 1 shows for stores who opt for promo. Those store who took promotion their sales are high as compared to store who didn't took promo

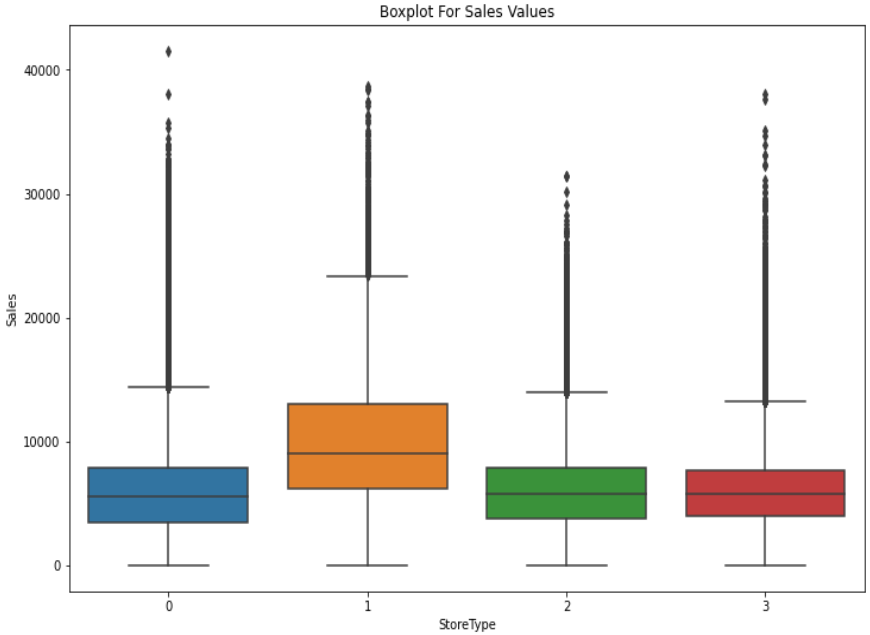
State Holiday & School Holiday

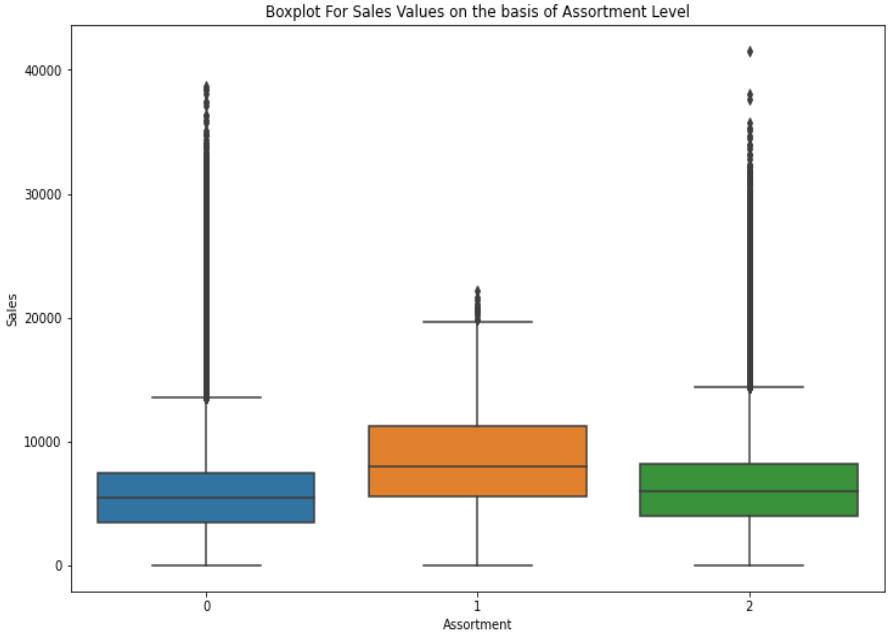




- We can observe that most of the store remain closed during state and holiday. but it is interesting to note that the no of store opened during school holiday were more than that were opened during state holiday. Another important thing to note is that the store which were opened during school holiday had more sales than normal.

Box plot of sales b/w assortment and store type





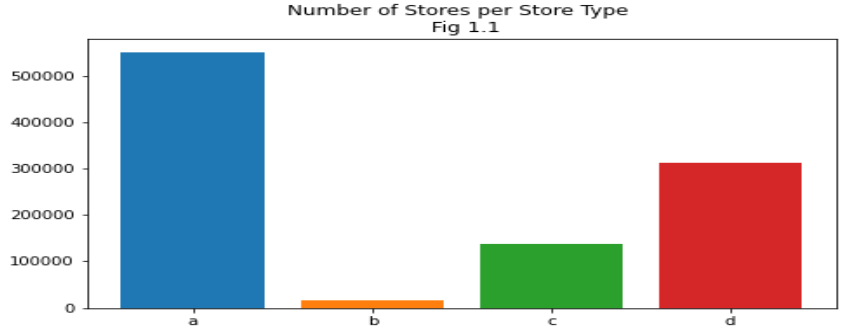
As we cited in the description, assortments have three type and each store have defined type and assortment type

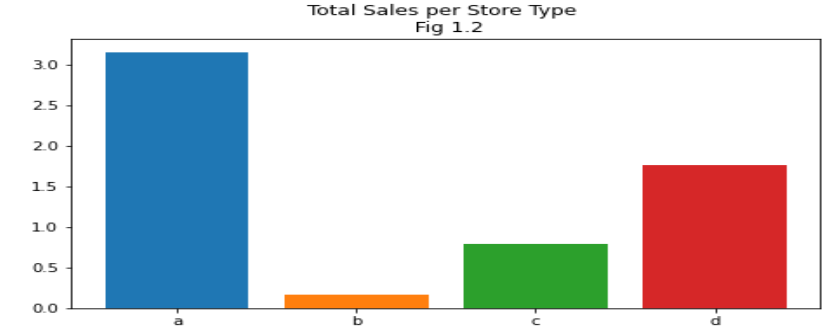
1) a means basic things

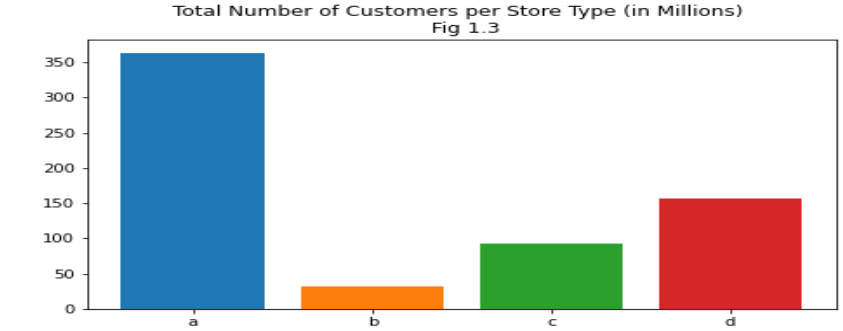
2) b means extra things

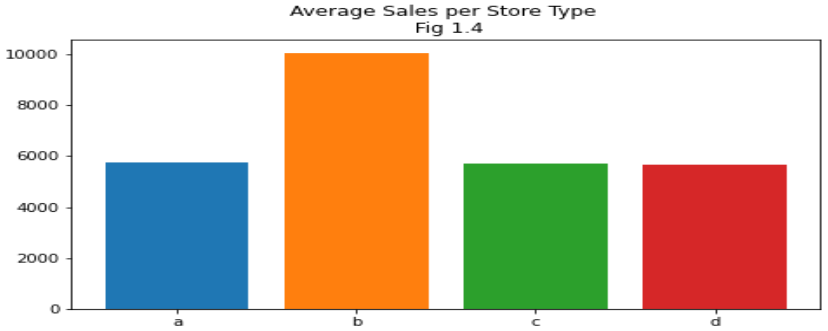
3) c means extended things so the highest variety of products

Store Type





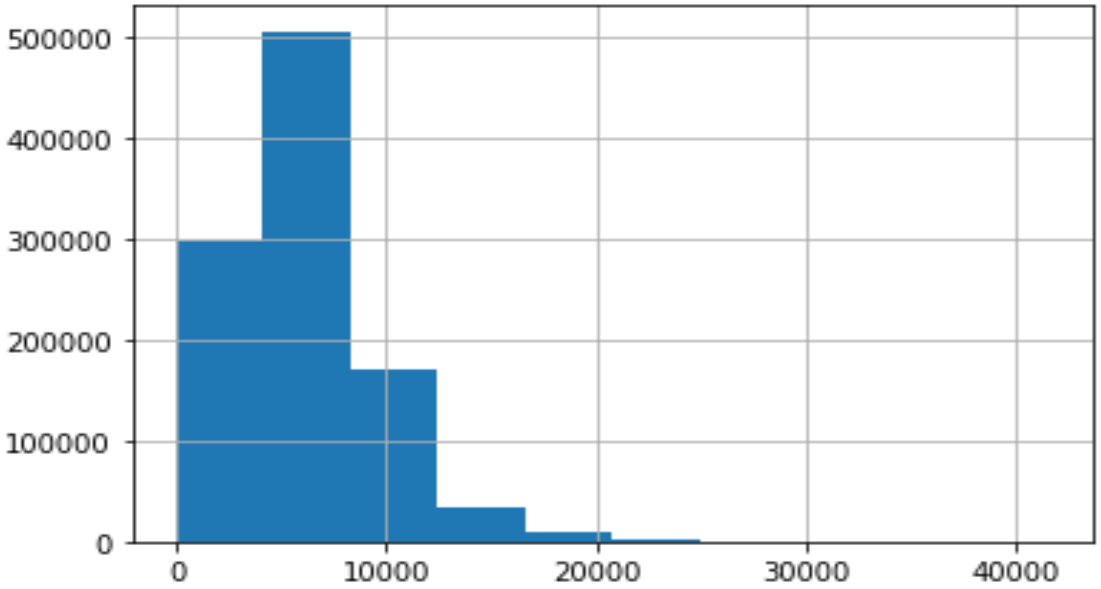




we can see that storetype A has the highest number of branches, sales & customers from the four different storetype but it doesn't mean its the best performing storetype

when looking at the average sales and number of customers, we see that actually it is a storetype B who was the highest average sales and the highest average number of customers.

Sales Dependent Variable



- Now checking the number of sales =0

Machine learning

- Machine learning is a type of artificial intelligent that allow software application to become more accurate at predicting outcome without being explicitly programmed to do so machine learning algorithms use historical data as input to predict new output value
- Machine learning are a common use case for machine learning other popular uses include fraud detection, business process automation (BPA) and predictive maintenance

Why machine learning is important

- Machine learning is important because it gives enterprises a view of trends in customers behavior and business operational patterns , as well as supports the development of new products. Many of today leading companies such as Facebook, Google and Uber, make machine learning a central part of their operation. Machine learning has become a significant competitive differentiator for many companies

Different types of machine learning

- Supervised learning : In this types of machine learning, data scientists supply algorithms with labeled training data and defined the variables they wants the algorithms to assess for correlation.
- Unsupervised learning : This types of machine learning involves algorithms that train on unlabeled data.
- Reinforcement learning: It is multi step process for which there are clearly defined rules Data scientist program algorithm complete a task

Linear Regression

Linear Regression is a kind of parametric regression model that makes a prediction by taking the weighted average of the input features of an observation or data point.

- **Advantage and disadvantage of linear regression:**
- Linear regression is a simple to implement , and on other hand in linear regression in difficult
- The relationship between the independent variable have a linear relationship, and other hand linear regression also looks at a relationship between the mean of the dependent variable and the independent variables .

Import Linear Regression:-

```
from sklearn.linear_model import LinearRegression
```

Decision tree

Decision Tree is the most powerful and popular tool for classification and prediction. Decision tree is a flowchart like tree structure, where each internal node denotes a test on a attributes, each branch represents an outcome of the test, and each leaf node(terminal node) holds a class label.

Advantages and disadvantage of the decision tree:

- They are very fast and efficient compared to KNN and other classification algorithms. Each to understand, Interpret, visualize. The data type of decision tree can handle any type of data whether it is numerical, categorical or Boolean
- One of the limitations of decision tree is that they are largely unstable compared to other decision predictors . A small change in the data can result in a major changes in user will get in a normal event.

Import Decision Tree regressor:-

```
from sklearn.tree import DecisionTreeRegressor
```

Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Advantages of Random Forest:-

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over fitting issue.

Disadvantages of Random Forest:-

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

Import Random Forest Regressor:-

- from sklearn.ensemble import RandomForestRegressor

Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage shrinkage is where data value are shrunk towards a central point.

Advantage and Disadvantage of Lasso Regression:

The main advantage of a lasso regression model is that it has the ability to set the coefficient for feature it does not consider interesting to zero and the model does some automatic feature selection on other side biased coefficient that are produced by a lasso model are biased. The L1 penalty that is added to the model artificial shrinks the coefficient closer to zero and difficult to estimate standard errors since the coefficient estimates in a lasso model are biased.

Import LASSO

```
from sklearn import linear_model.Lasso
```

Ridge Regression

- Ridge regression is a model tuning method that is used to analyze any data that suffers from Multicollinearity . This method performs regularization

- **Advantage and Disadvantage of Ridge regression:**
- The biggest benefit of ridge regression is its ability to produce a lower test mean squared error (MSE) compared to its least square regression when multicollinearity is present
- The biggest drawback of ridge regression is its inability to perform variable selection since it include all predictor variable in the final model since some predictors will get shrunken very close to zero.

Import RIDGE

```
from sklearn import linear_model.Ridge
```

Challenges

- Understand the column of the dataset.
- Analyze and visualization of the rossmann sales according to feature.
- Find the right chart to show the chart.
- Difficulties to find out the correct model.

Conclusion

- From plot sales and competition open since months show sales go increasing from November and highest in months December.
- From plot sales and day of the week sales highest on Monday and start declining from Tuesday to Saturday and o Sunday sales almost near to zero.
- Plot b/t promotion and sales show that promotion helps in increasing sales.
- Type of store play an important roles in opening pattern of stores.
- All Type'b'stores never closed except for refurbishment or other reason.
- Assortment level 'b' is only offered at store Type 'b'.
- We can observe that most of the stores remain closed during state Holiday. But it is interesting to note that the no of store opened during school holiday.

Q/A