**Retail Sales Prediction**

**Mohan Vishe, Shambhuraj Desai,**

**Gaurav Yadav, Rahul Ray**

**Data science trainees,**

**AlmaBetter, Bangalore**

**GitHub link**

**https://github.com/MohanVishe/Supervised-ML--Regression----Retail-Sales-Prediction**

**Google Drive**

**https://drive.google.com/drive/folders/1Ezu6B68vvjh922sTeKIhgeYLEX7v8OPT?usp=share\_link**

**Individual colab project**

**https://colab.research.google.com/drive/1W7VaAIH38Lp3mtXmT2yraBaejBWG\_l67?usp=share\_link**

**Team colab project**

**https://colab.research.google.com/drive/18UNkvchbr4lSPJOBXoLaXMJB36jKPUs2?usp=share\_link**

**Abstract:**

Client interest and demand for product changes with time. Which is important for growing business. 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 Sales prediction is estimating the demand for a service or product for a particular time. This project contains the real-world business problem of sales prediction with the help of machine learning models.

Our task is to predict sales and find important factors affecting demand to improve sales.

Here we are predicting sales of Major store chain Rossmann

**Keywords: EDA, Transformation, Preprocessing, Correlation, Decision Tree Random Forest, Regression, Forecasting**

**1. Problem Statement**

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales 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

There are so many features present in the dataset which contain various keywords.

Data fields  
Most of the fields are self-explanatory. The following are descriptions for those that aren't.  
  
\* Data fields  
Most of the fields are self-explanatory. The following are descriptions for those that aren't.

1. Id - an Id that represents a (Store, Date) duple within the test set  
2. Store - a unique Id for each store  
3. Sales - the turnover for any given day (this is what you are predicting)  
4. Customers - the number of customers on a given day  
5. Open - an indicator for whether the store was open: 0 = closed, 1 = open  
6. 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  
7. School holiday - indicates if the (Store, Date) was affected by the closure of public schools  
8. StoreType - differentiates between 4 different store models: a, b, c, d  
Assortment - describes an assortment level: a = basic, b = extra, c = extended  
9. CompetitionDistance - the distance in meters to the nearest competitor store  
10. CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened  
11. Promo - indicates whether a store is running a promo on that day  
12. Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating  
13. Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2  
14. 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, and November of any given year for that store

**2. Introduction**

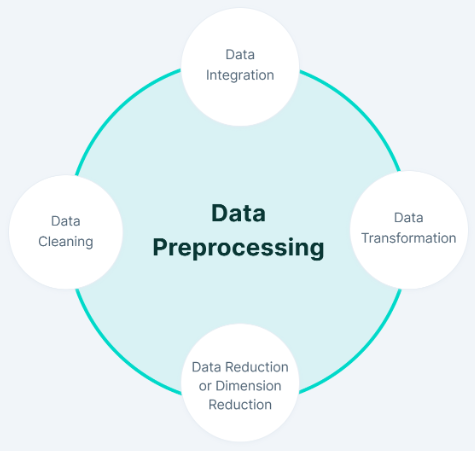
To increase the business we should focus on client interests and demand for products in future

important. Sales forecast helps

To predict those trends and demands, we required good models which would be trained on a good dataset.

Take a look at the short map of our journey..!

i) Loading the dataset



ii) Cleaning the dataset

and prioritizing our main objectives.

iii) Using some techniques we set our data and merge data\

iv) Hypothesis

v) Using some techniques we do statistical analysis.

v) Conclusion.

vi)Features transformation

vii) Data splitting

viii) Model Training

## **4. Hypothesis**

1. Promotion will leads to an increase in sales  
2. At the weekend the sales would be low as stores will be closed  
3. Holidays will lead to a decrease in sales  
4. The number of customers will positively correlate with sales  
5. Sales will be zero when stores are closed

**5. Steps involved:**

**i) Loading the dataset**:

There are 4 libraries used which arenumpy, pandas, matplotlib.pyplot, and seaborn. Pandas is a great library for EDA. Here Data is in '.csv' format. We use the panda's library to load the dataset into the notebook, after mounting the drive and putting the path of the dataset we then read the data." pandas.read\_csv() to convert it into a data frame. This method takes the path of a CSV file. To find the number of rows and columns of the data we use ".shape" where ".info()" gives information about columns. We have two data set of shape(1017209, 9), and (1115, 10) dataset we have to do some cleaning.

**ii) Cleaning the dataset:**

The cleaning process involves removing the dataset values which are unnecessary for the objectives, also it includes some columns which have some null values. We transform the dataset into a consistent format to ensure we predict some good results.

Null values are the data that cannot provide any information and create obstacles to reach out the goal. Data Pre-processing is important before doing EDA as there are many outliers and missing values present in the dataset. We use ".isna().sum()" to find missing values in the dataset. The dataset contains missing values in a few columns.

To save data there are some NaN values that are replaced by mean, median or mode

Convert the Date feature to date and time format.

a median, or zero. To replace these

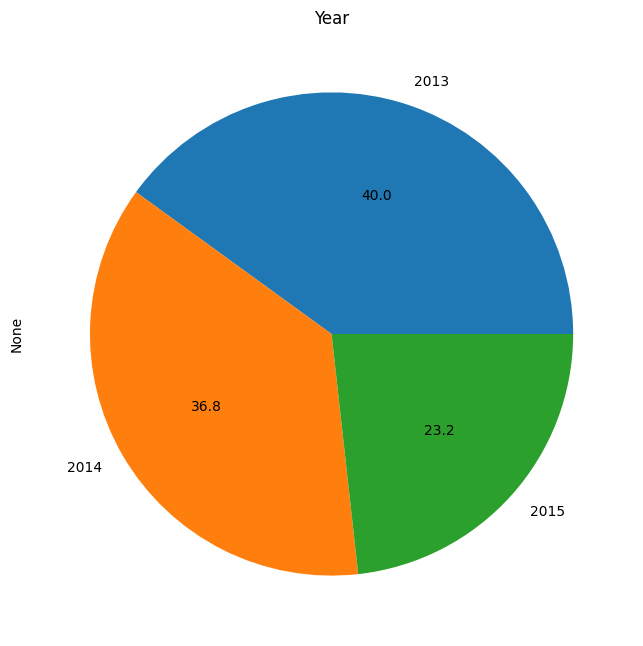
values we use ". fillna()". There

are some values converted into

integers where required.

.

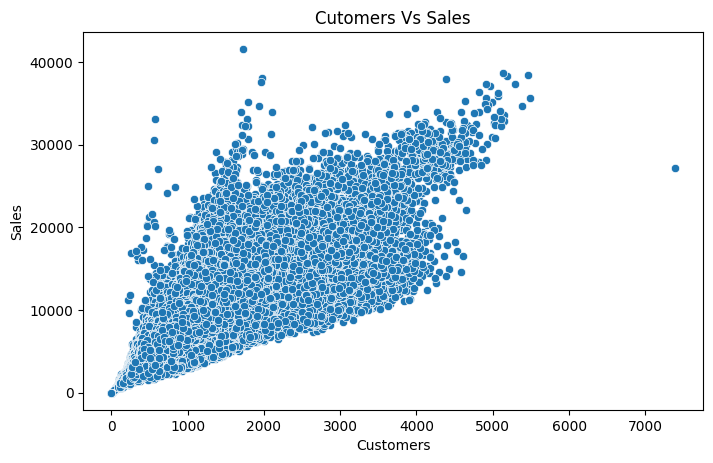
**6. Exploratory Data Analysis**



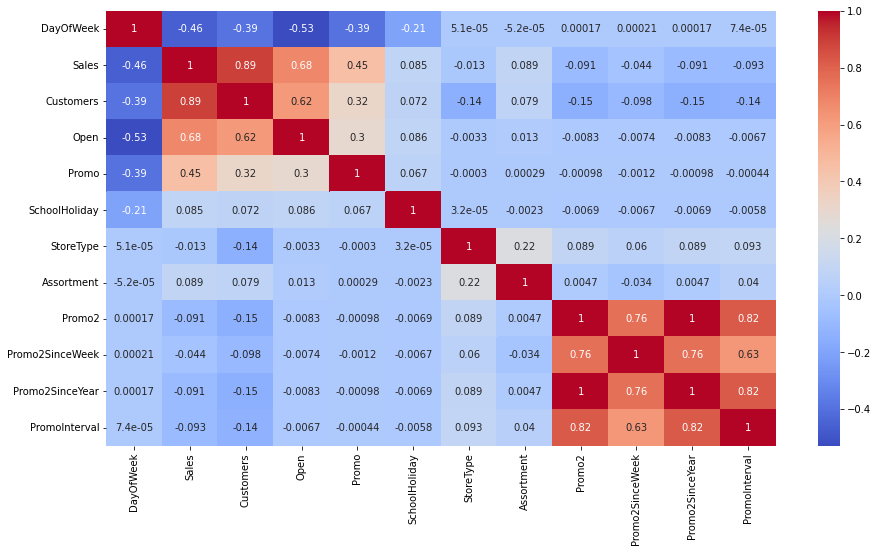
**a) Value Counts:**

38.2 % of data has promotion involved.  
54.2% of data is of store type "a".  
17% of data is of closed stores.  
Data contain 2013 values around  40%

**d) Scatter Plot**

The competitor stores are close to each other has more sales. We can see that sales are positively correlated with customers

**c) Correlation Between Features**



for visualization of correlation

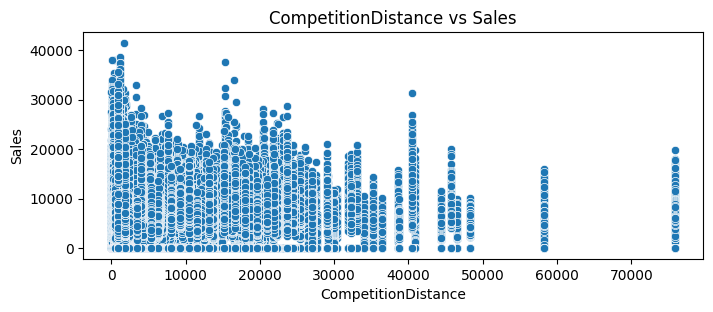
between features heatmap is

the best option that we have.

Customers, Sales, Open, and Promo are high and positively s correlated with each other which is understandable.

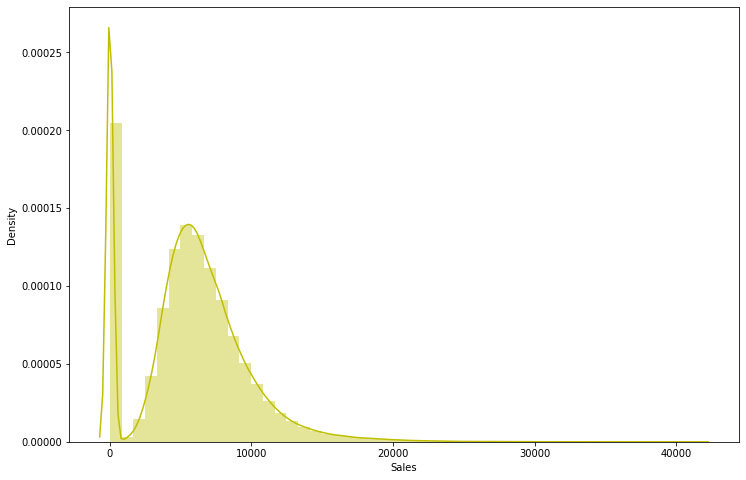
Whereas DayOFWeek has a negative correlation with these features.

'Promo2', 'Promo2SinceWeek', and 'Promo2SinceYear' has some correlation with each other



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.

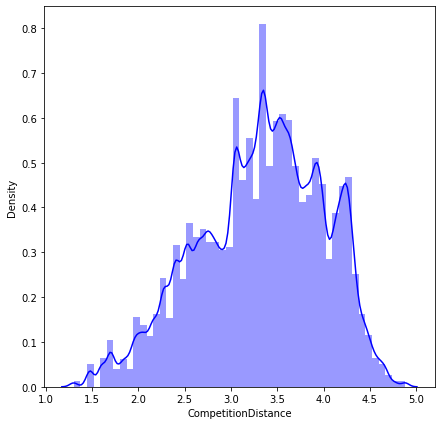
**e) Bar Plot / PDF**



There are some sales values which are shown as zero this is

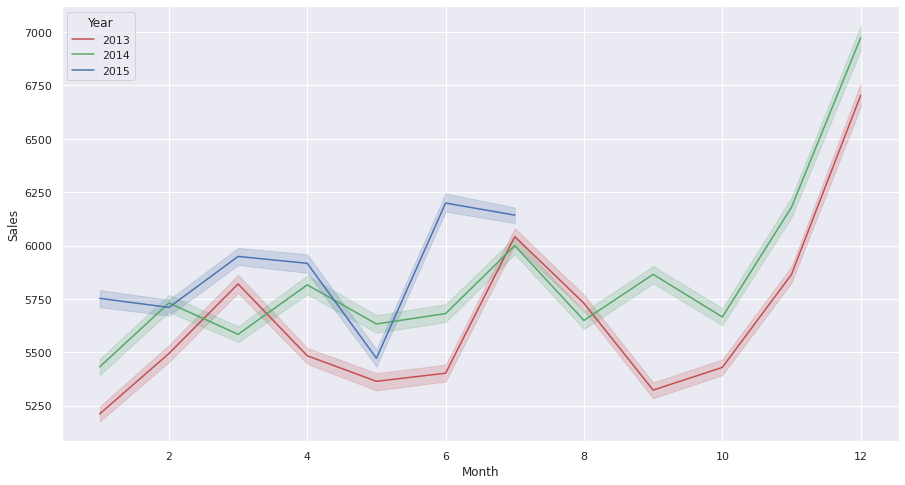
because the stores were closed due to refurbishment

**f) Log transformation**



After applying log transformation on the “CompetitionDistance” the skewness is reduced this will also reduce the impact of outliers.

**g) The yearly comparison:**



At end of the month the sales are much higher compared to other months.

**7 Feature Transformation:**



Normalization: is a scaling technique in which values are shifted and rescaled between 0 and 1



Standardization: is another scaling technique where the values are centred around the mean with a unit standard and deviation. The range is [-1,1]

**One hot encoding:**

Here we convert categorical data into binary features with

value {0,1} we have done one

hot encoding on features as DayofWeek, StoreType and Assortments

Decision Tree is a Supervised learning technique that can be used for both Classification and Regression problems

**Train-Test Split:**

we randomly splitting data into

train and test data where 80 per cent of data used for training and cross-validation and remaining data used for testing

**8. factors affecting choosing the model**

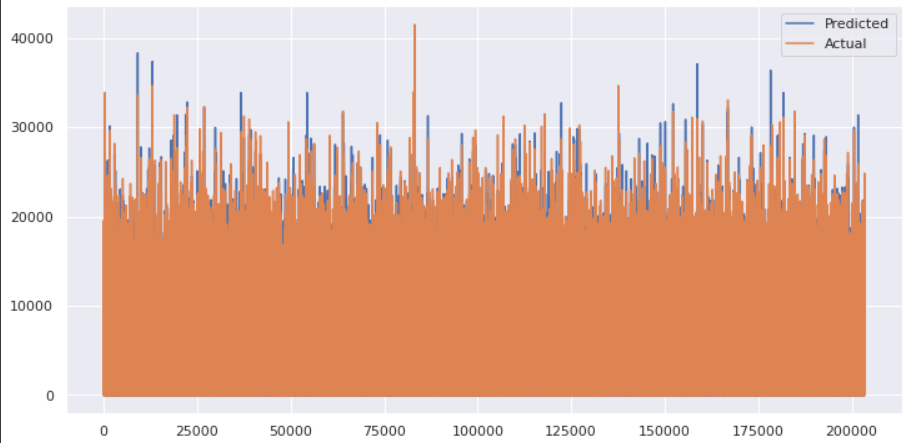
1. There are various factors such as time complexity, categorical features, numerical features, data, outliers, and noise which decide the best model to be used.  
2. As there are many categorical features so the decision tree could work better in such a case  
3. decision Tree is highly interpretable which will helpful for feature importance  
4. Here we go with linear regression models to understand which will work better.

**9. linear Regression:**

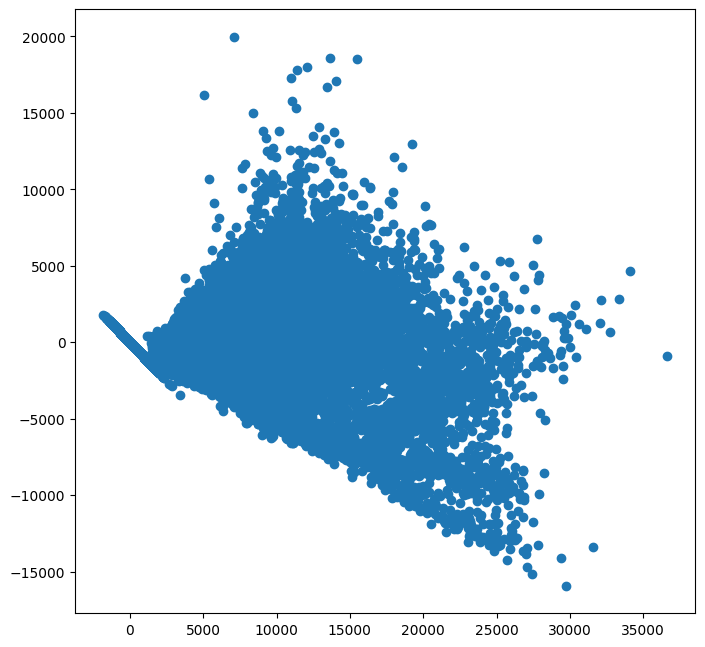
Mode performance is not bad.

Because a large part of the data is categorical it doesn’t perform as

well as other models



Heteroscadacity



we also applied rigid and lasso regression which also gives us similar performance.

**10) Decision Tree, Random forest:**

It gives us better performance than linear regression because it is also good at regression tasks and it handles categorical

features very well.

Where the Random forest is the best model we get after training

**11) Time Splitting:**

We divide the dataset into two parts training and testing. Where the data for the last 6 weeks comes in the testing and the remaining data in training.

As we have to predict data from the last 6 weeks means we have to train on data before the 6 weeks

In such cases random split will lead to data leakage

So we go with time base split for our best model.

**12) Cross Validation:**

We find the best hyperparameter which gives us the best result. We use grid search for this task

There are slightly better results by using a good hyperparameter.



**13. Observation:**  
  
1. From plot sales and competition Open Since Month shows sales go increasing from November and highest in month December.  
2. From the plot Sales and day of the week, Sales are highest on Monday and start declining from Tuesday to Saturday and on Sunday Sales are almost near Zero.  
3. Plot between Promotion and Sales shows that promotion helps in increasing Sales.  
4. Type of Store plays an important role in the opening pattern of stores.  
5. All Type ‘b’ stores never closed except for refurbishment or other reasons.  
6. All Type ‘b’ stores have comparatively higher sales and it is mostly constant with peaks appearing on weekends.  
7. sort ment Level ‘b’ is only offered at Store Type ‘b’.

**14. Conclusion:**

1. Rossmann Stores Data.csv dataset has 10,107,219 rows and 10 columns. store.csv dataset has 1115 rows and 9 columns.  
2. The sales in the month of December are the highest sales among others.  
3. The Promotion increases sales so we should focus on that factor  
4. As the customers are positively correlated with sales so we have to increase the frequency of customers by offers  
5. The sales for store type B are higher than any other stores

**References**

1. pandas.pydata.org

2. GeeksforGeeks

3. Seaborn.pydata.org

4. Matplotlib.org

5. Numpy

6. Scikit-Learn Org

7. Kaggle

8. Towards Data Science Blogs