Rossmann Sales Prediction

Satyam jyoti sankar

Data science trainees,

AlmaBetter, Bangalore

Abstract :

Rossmann is one of the largest drug store chains in Europe operates over 3,000 drug stores in 7 European countries as for our dataset . Rossmann store managers are take care of estimate how sales going on 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 in case to case. Our main task is to do same thing through machine. The goal is to forecast the "Sales" for the test set. So that we conclude how machine performing in compare to the managers.

. With the help of the Exploratory Data Analysis on the given data set we finding the correlation of the various feature affected the sales. The objective of the model building from the given data we perform the various regression technique to get the best result for that we perform the train test split with the important feature. This gives us the good predictive module to reach out our main goal of predicting sales and se how it work on compare to managers prediction.

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.

I have an historical sales data for 1,115 Rossmann stores.

The task is to forecast the "Sales" column for the test set.

Data set Description:

Here we have 2 data set:-

Rossmann Stores Data.csv - historical data including Sales.

store.csv - supplemental information about the stores.

Data fields:

- Id - an Id that represents a (Store, Date) duple within the test set
- Store - a unique Id for each store
- Sales - the turnover for any given day (this is what you are predicting)

- Customers - the number of customers on a given day
- Open - an indicator for whether the store was open: 0 = closed, 1 = open
- State Holiday - 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
- School Holiday - indicates if the (Store, Date) was affected by the closure of public schools
- Store Type - differentiates between 4 different store models: a, b, c, d
- Assortment - describes an assortment level: a = basic, b = extra, c = extended
- Competition Distance - distance in meters to the nearest competitor store
- Competition Open Since [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

1. Factors Affecting:

Following are the factors affected sales more :

1. **Customer.:** Our target variable is sales when we compare with no of customer here we see when no of customer increase no of sales also increase..

2. **Day of week.**: High sales in starting of week. On Sunday only some stores are open and they are making great Sale.

3. **State Holiday.**: There are very few stores open on 'State Holiday' and they make a good profit on those days then any average day.

4. **School Holidays.**: On School Holidays there is no large difference in sale. So promos running on School holidays can be reduced.

5. **Assortment.:** Sales for assortment type a and c seems to be less as compared to assortment type b.

6. **Competition Open Since [Month/Year]:** At the start of month the sales increases. In compare of other.

Steps involved :

The following steps are involved in the project

1. **Exploratory Data Analysis** :
   After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is sales with other independent variables.
   This process helped us figuring out various aspects and relationships among the target and the independent variables and also we observed the distribution of variables.
   It gave us a better idea that how feature behaves with the target variable.

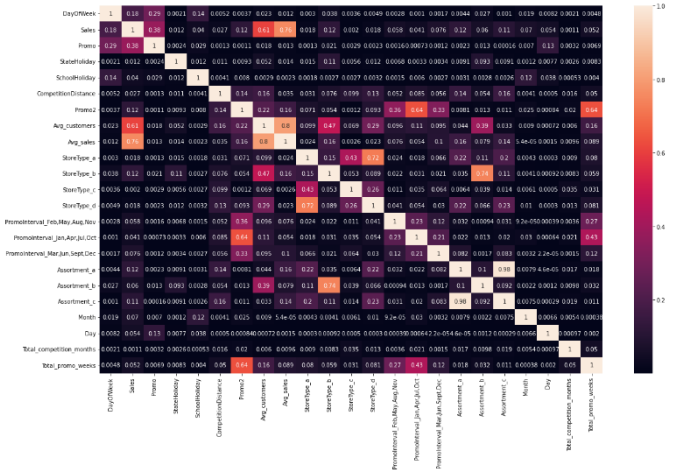
2. **Preprocessing data** :
   The fast Dataset contains a no null values also noduplicate values but in 2nd data set contain some missing value.
   Dropping the unwanted columns and adjust the missing value I combine 2 data set and create a single final dataset.
   In final data set I have 840211 rows and 26 columns.

3. **Features selection.** :

In features selection we take selected features and take them to final dataset.
In final data set we have 26 features among them we combine some features and the end for model implementation we have 23 features dataset and sales is my target value.

4. Correlation Analysis :

For correlation analysis visualization we use heat map it show correlation between dependent and independent variable



On base of correlation I adjust the features like if they are highly corelated combine them and make a single variable.

5.Train test Split :

In train test split we take 'x' as dependent variables and 'y' take as independent variable then train the model.

y show no of sales

6.Models :

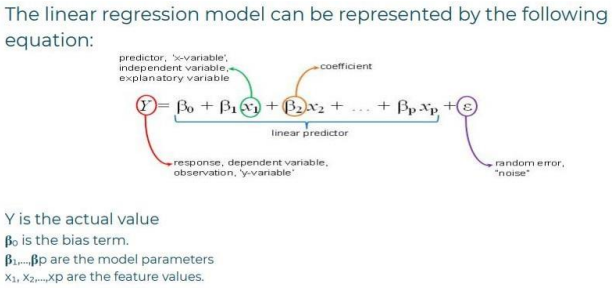
Here I use 6 model to train the data andfor predicting the accuracy R2 and adjR2.

1. Linear Regression (Baseline Model)
2. Lasso (Hyperparameter)
3. Decision tree by using decision tree regressor (Hyperparameter)
4. Random forest regressor
5. Gradient Boosting Regression
6. Xg boost regressor (Hyperparameter)

Modeling Results.:

1.Linear Regression (Baseline Model):

The hypothesis of linear regression model represented below:



We have model by linear regression and we get results as follows:

Score on test data.:

MSE.: 72.40464240122799
RMSE.: 8.509091749489365
R2.: 0.7438439905259265
Adjusted R2 :0.7438216313375

2. Lasso (Hyperparameter)

By performing lasso regression (hyperparameter) we get the results are as follow.:

Score on test data:

MSE.: 72.40978143885137
RMSE.: 8.509393717466091
R2.: 0.7438258094352328
Adjusted R2.: 0.74380344865983

It gives similar output like our base model

3. Decision tree.:

By performing Decision tree we get the are as follows.:

Score on test data.:

MSE.: 59.419466591675246
RMSE.: 7.708402337169178
R2.: 0.7897834594243723
Adjusted R2.: 0.7897651101720973

4. Random forest regressor.:

By performing l Random forest regressor we get the following result.

Score on test data.:

MSE.: 20.035760167057393
RMSE.: 4.476132277654157
R2.: 0.9291166947178242
Adjusted R2.: 0.9291105074994105

It gives very good score.

5. Gradient Boosting Regression.:

By performing Gradient Boosting Regression we get the following result.

Score On test data.:

MSE.: 46.62906371819621
RMSE.: 6.8285477019785255
R2.: 0.8350338529209714
Adjusted R2.: 0.8350194534572503

6.Xg boost Regression (Hyperparameter)

By performing XG Boosting Regression

we get the following result.

Score On test data.:

MSE.: 27.615515282849135
RMSE.: 5.255046648969839
R2.: 0.9023007370822268
Adjusted R2.: 0.902292209168973

Conclusions.:

- Random forest regressor gives us high accuracy of 93% for our test data set in case of train data it show accuracy of 99% .

- Xg boost gives 90% in both train and test case both r2 score and adj r2.

- In case of random forest a little bit overfit occur but in xg boost it work properly.

- Our base model liner regression created a base line of accuracy of 74%.

- The lasso model although we use hyper parameter it gives us similar performance like linear regression.

- In case of Decision tree by using decision tree regressor (Hyperparameter) it gives us 78% accuracy in train and 79 accuracy for test case

- Random forest and xg boost work good on our sales prediction, both are gives more than 90% accuracy in prediction on test data.