RETAIL SALES PREDICTION

SUPERVISED MACHINE LEARNING - REGRESSION

Baliram Kumar

AlmaBetter, Bangalore(https://www.almabetter.com/)

Abstract: In the following project, we have applied machine learning to a real world problem of predicting retail stores sales. Such predictions help store managers in creating effective staff schedules that increase productivity. We used the popular open source programming language Python and used its libraries like NumPy, scikit-learn, pandas , matplotlib for modelling, analysis and prediction and visualization. We used feature selection, model selection to improve our prediction result. In view of the nature of our problem, Root Mean Square Error (RMSE) is used to measure the prediction accuracy.

Keywords: Sales Prediction, NumPy, scikit-learn, machine-learning, RMSE, Linear regression, lasso regression ,decision tree

1. Introduction: In any supply chain, an ability to accurately predict sales has a direct impact on its operating expenditure. Being able to accurately predict the sales validates understanding of the factors influencing it. A good understanding of these underling factors enable in taking "decisions" that can improve sales.

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

2. Dataset and Features

Training data consists of two parts. One part is historical daily sales data of each store from 01/01/2013 to 07/31/2015. This part of data has about 1 million entries. Data included multiple features that could impact sales. Table 1 describes all the fields in this training data.

Table 1: Historical sales data table features

Field Name	Description
Store	a unique Id for each store: integer number
DayofWeek	the date in a week: 1-7
Date	in format YYYY-MM-DD
Sales	the turnover for any given day: integer number (This is what to be predict)
Customers	the number of customers on a given day: integer number (this is not a feature. Based on the data set from AlmaBetter site, this feature is not included in dataset)
Open	an indicator for whether the store was open: 0 = closed, 1 = open
Promo	indicates whether a store is running a promo on that day: 0 = no promo, 1 = promo
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: 1 = school holiday, 0 = not school holiday

The second part of training data is supplement store information. It has 1115 store info entries, which listed the store type, competitor and a different kind of promotion info. Table 2 below describes all the fields in this file.

Table 2: Store Information data table features

Field Name	Description
Store	a unique Id for each store: integer number
StoreType	differentiates between 4 different store models: a, b, c, d
Assortment	describes an assortment level: a = basic, b = extra, c = extended
CompetitionDistance	distance in meters to the nearest competitor store
CompetitionOpenSinceMonth	gives the approximate year and month of the time the nearest competitor was opened
CompetitionOpenSinceYear	
Promo2	Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
Promo2SinceWeek	describes the year and calendar week when the store started participating in Promo2
Promo2SinceYear	
Promo Interval	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

Dataset Statistic:

STATISTICS	NUMBERS
Rossmann Dataset size	1017209
Total stores number	1115
Rossmann data Time ranges	2013-08-01 to 2015-07-01

We did several things to combine features and create features directly related to sales numbers. The work we did is:

1. The supplement store information can’t be used directly. We merged store information and historical sales data. Store type and Assortment is merged into each entry of historical sales data

2. Combine Promo2, Promo2SinceWeek, Promo2SinceYear and Promointerval to a promotion 2 indicator in historical sales data. The indicator indicates on a certain day whether a certain store is on promotion 2.

3. Similarly, we combined CompetitionDistance, CompetitionOpenSinceMonth, CompetitionOpenSinceYear to a competitor indicator. The indicator indicates on a certain day whether a certain store has a competitor.

4. Since CompetitionDistance is provided, we used CompetitionDistance to train the model, instead of competitor indicator. For any date and any store which doesn't have a competitor (competitor==1), we assign CompetitionDistance as a large number 100000. This method enables us to use only one CompetitionDistance feature. It also models the no competitor case by weakening CompetitionDistance impact.

5. Historical sales dataset has Date feature. We created a Month and Year feature based on the Date feature. Month and Year are used as features, since they correlate with sales data.

The final training dataset used includes the following features.

StoreID	Open	Promo2 indicator
DayOfWeek	StateHoliday	Store Type
Month	SchoolHoliday	Assortment
Year	Promo	CompetitionDistance

3. Literature review:

Forecasting is projecting, predicting or estimating some future condition or event that is beyond an organization's power and gives a basis for efficient planning. Forecasting is necessary for several situations of modern business and its proper working. Organizations must make plans that will be efficient at some point in the future. And to do this they require information and data about current circumstances. It is very unfortunate that though forecasting is an important aspect yet its progress in many fields or research and development has been limited. In the past decade Machine Learning has emerged as a technology with a great promise for identifying and modeling data patterns that are not easily described by traditional statistical methods in a field as diverse as cognitive science, computer science, electrical engineering and finance

- What is the extent to which sales performance is influenced by factors like: promos, school and state holidays, competition distance ,competition open month. locality and seasonality,

- What model is appropriate to predict sales?

4. Problem Formulation:

Rossmann store managers had to predict the daily sales and the number of customers for up to six weeks in advance; while store sales, What is the extent to which sales performance is influenced by factors like: promos, school and state holidays, competition distance ,competition open month. locality and seasonality, As there are so many individuals who try to forecast sales based on their unique sets of circumstances, the accuracy of such forecasts was rather varied. So our task was to make an efficient machine learning model that would predict the sales for 1,115 stores across Germany using which store managers would be able to create effective staff schedules to increase their productivity and sales turnover.

We are given a dataset from Rossman inc. An algorithm that will predict the quantity of sales. The evaluation metric for this problem is RMSE. The RMSE calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

5. Future scope of the project

Our model will help local retailers to spike their business in the following ways:-

1. It will help them to decide marketing strategies.
2. It will help them prepare the budget and for setting financial policies.

3. It helps in organizing stocks and prevents the risk of both overstocking and understocking.
4. With the help of forecasts we can find out which product provides more profit and which product’s manufactured should be stopped.

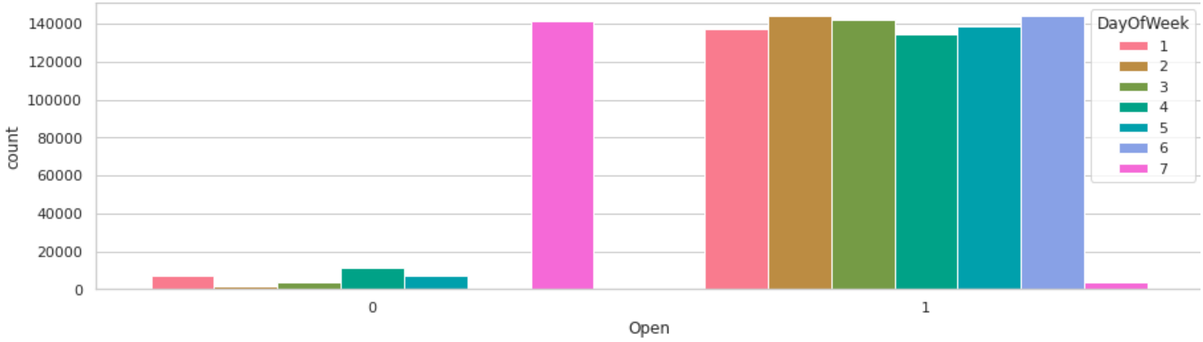
We believe every business will at some point in the future consider forecasting their sales for the upcoming challenges.

6. Analysis & Exploration:

In the following section we try to analyze our dataset and figure out what are the most important features for our predictive model out of useful features that can be used to forecast sales. At first we perform the feature extraction from our dataset and take out the derived values from the existing data given to us. Then, to get more important features, the store information is reviewed. At the end , we will try to figure out more information from store information.

Open

"Open" indicates if this store is open or not on a given specified day. Because the sales of the store must be 0 if it is closed, we removed the data point with (null value removal to reduce bias)"Open = 0" and after prediction, we will replace the value of sales as 0 for the data point with "Open = 0" in testing data.



It clearly shows that most of the stores remain closed during Sundays. Some stores were closed on weekdays too, this might be due to State Holidays as stores are generally closed during State Holidays and opened during School Holidays.

Year

The sales have a relationship with years, as the brand influence, marketing and other strategies of this company may vary from year to year, which could possibly have an impact on the sales.

Month

As people are more prone to having cold and other medical conditions during the winter and more sun related issues like dehydration and sunstroke during summer, people would have different demands for drugs during different months. So, the sales are possibly affected by the month of the particular year.

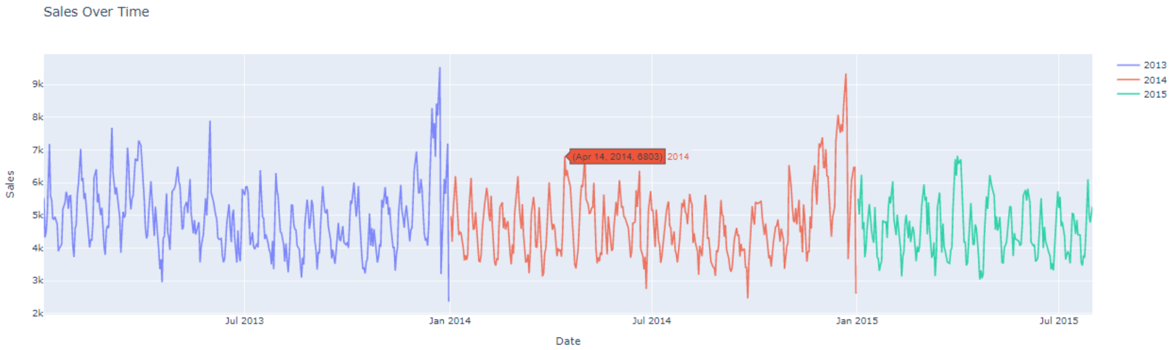
Day

Each single day could affect the sales. For instance, there may be people who tend to buy drugs on the first day of month or they might go to stores when they get their salary. So, days must also play a role in the sale pattern.

Store ID

Store ID is one unique feature as every store has its own different id’s. Sales may or may not change from store to store. If we use Store ID as a feature, we observe that the correlation coefficient of Store ID and Sales is 0.005.

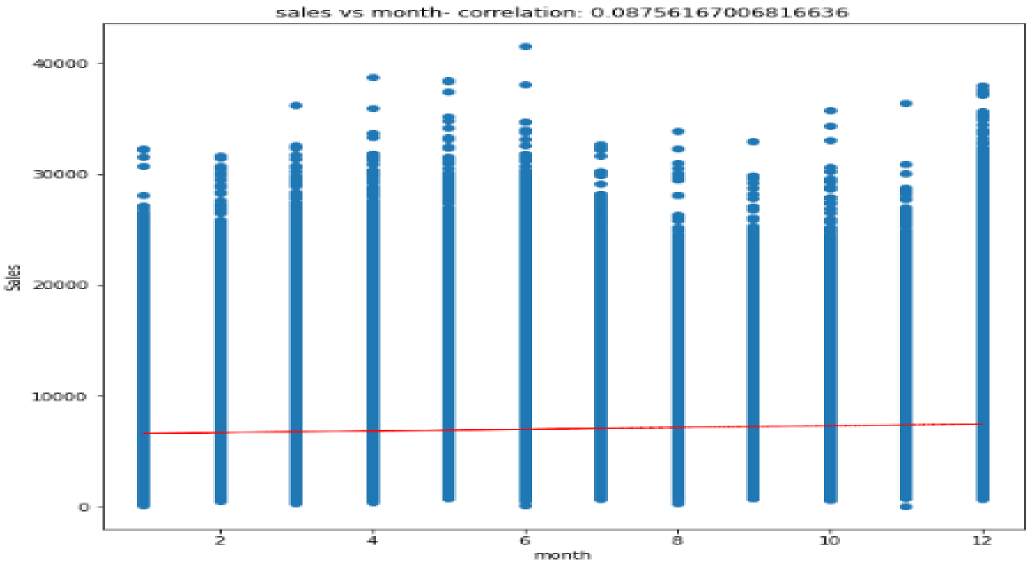
sales over time(year-month)

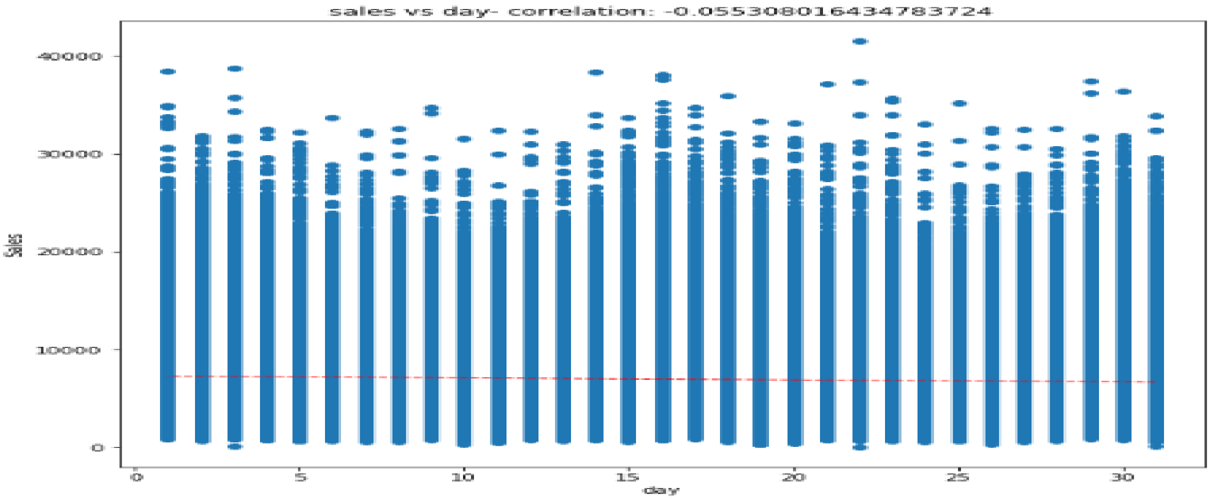


It is also interesting to note that Christmas and New Year(the above graph at weeks near 52) lead to increase in sales. As Rossmann Stores sells health and beauty products, it may be guessed that during Christmas and New Year people buy beauty products as they go out to celebrate and, this might be the cause of sudden increase in sales.

Sales on days and month

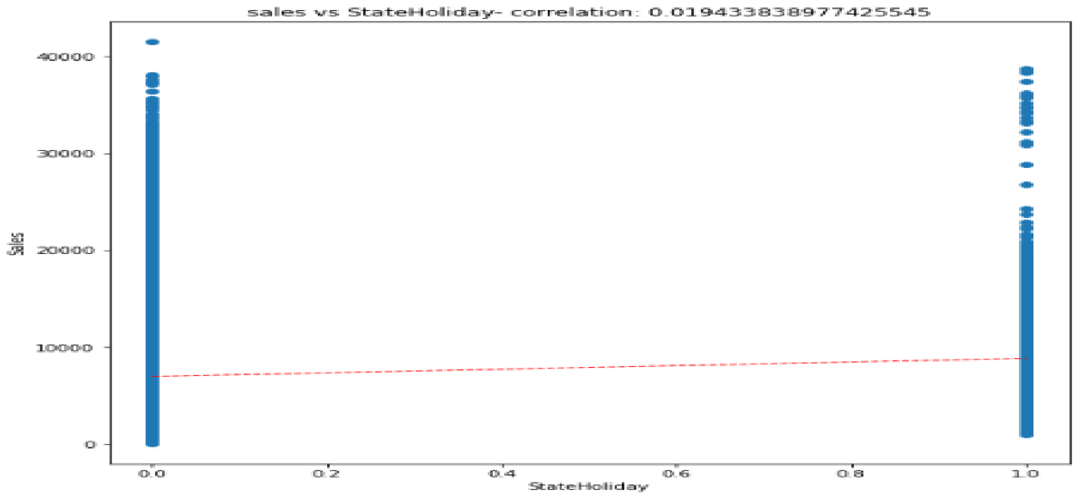
On different sales on days and month , each store will have different sales as people get used to shopping on different days. This particular feature plays a significant role in the sales prediction.





StateHoliday

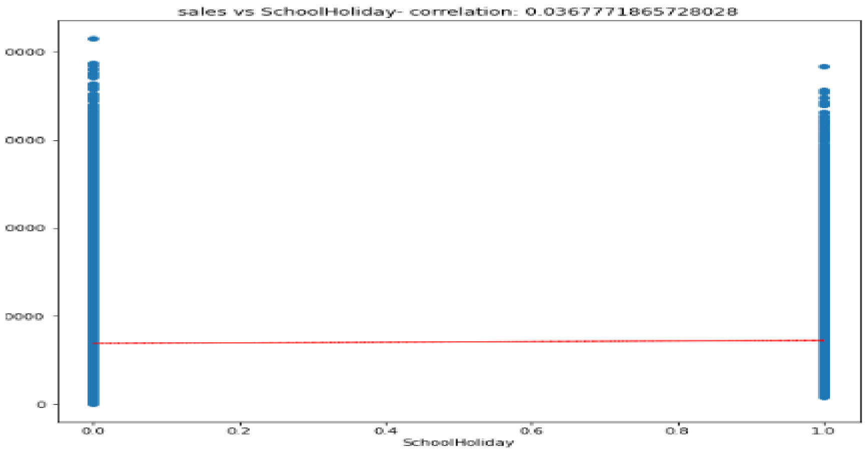
Different People have different demands and needs for drugs during holidays. We have information on state holiday, school holidays for each store every day. There is some correlation between state holidays and sales.



We have StateHoliday as Objects we need to convert them to numerical categories:
Which stores are closed in state holiday we are considers as 1 which days are not state holidays that days are considered as 0
Most of the stores remain closed during State and School Holidays. The number of stores opened during School Holidays were more than those opened during State Holidays. And the stores which were opened during School holidays had more sales than normal.

Sales affected by SchoolHoliday or not?

On examining the effect of school holiday on the effect of sales we can see the impact is not so significant.



We can observe that most of the stores remain closed during State and School Holidays. But it is interesting to note that the number of stores opened during School Holidays were more than those that were opened during State Holidays.

Another important thing to note is that the stores which were opened during School holidays had more sales than normal.

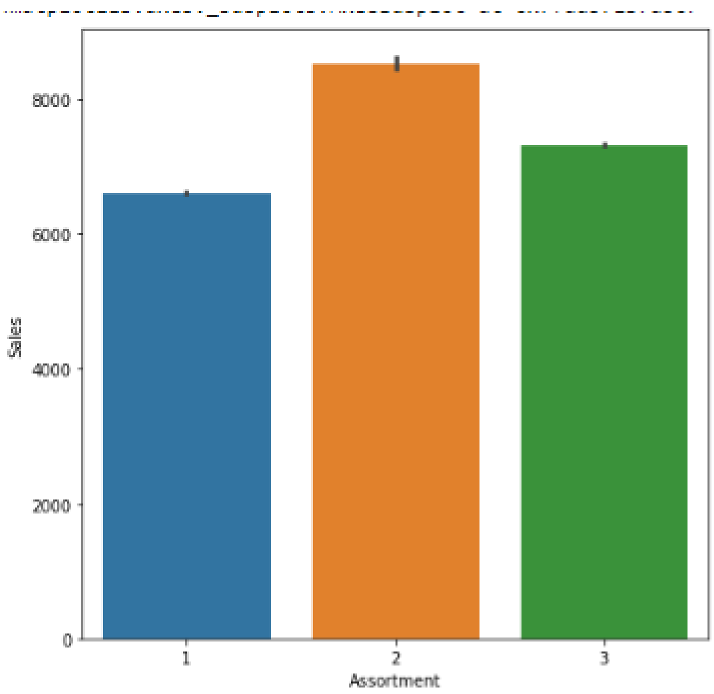
.

Assortment & Assortment Vs average sales and customers

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

- 1-means basic things
- 2-means extra things
- 3-means extended things so the highest variety of products.

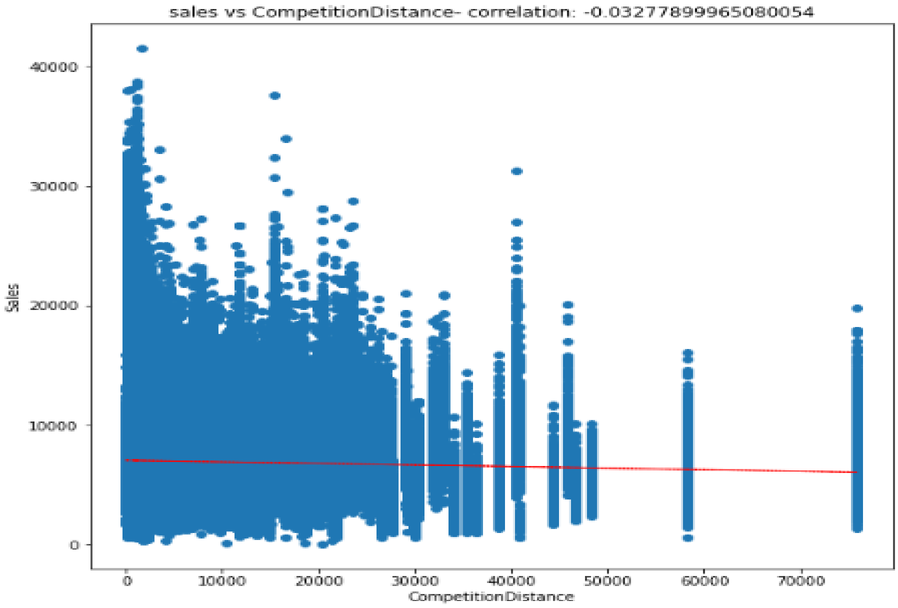
What could be interesting is to see the relationship between a store type and its respective assortment type.



We can clearly see here that most of the stores have either an assortment type or 3 assortment type. Interestingly enough, assortment type 2 has maximum sales and customers.

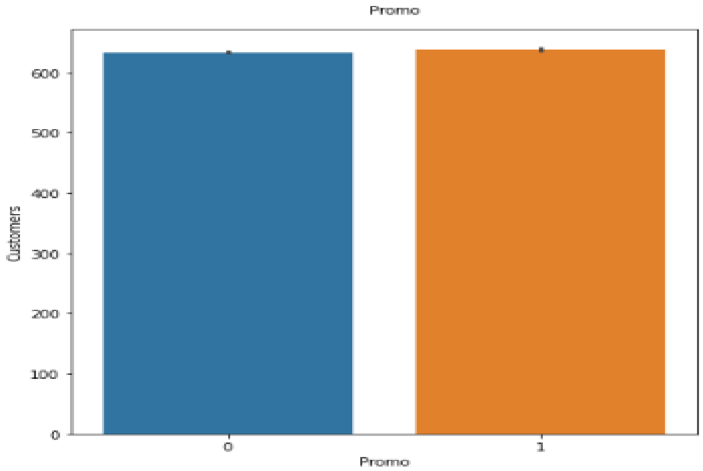
CompetitionDistance

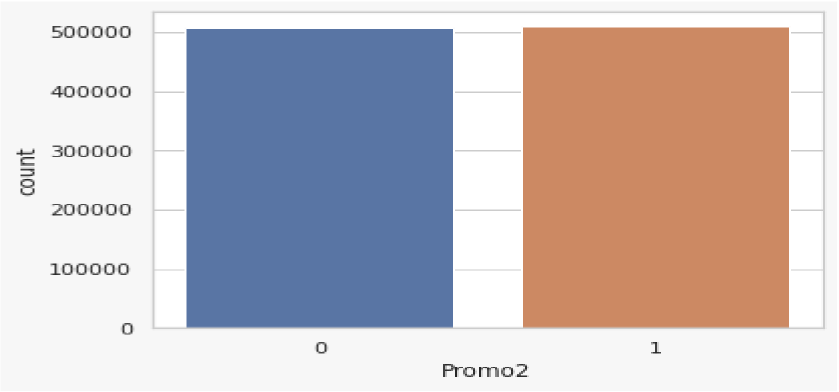
People will obviously prefer going to the store which is closer to their location. Their competition distance can also affect the sale of a store.



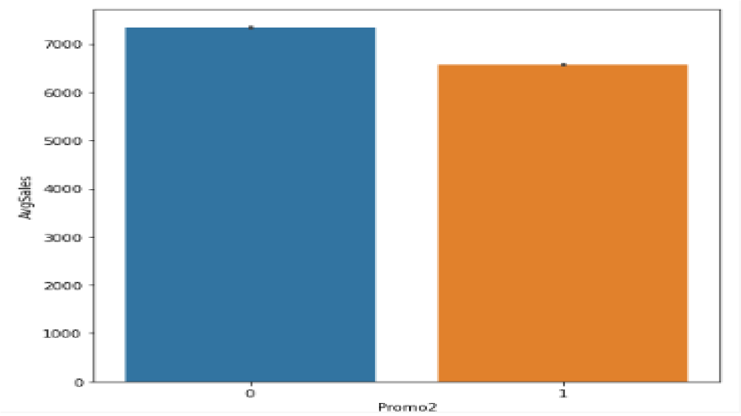
The stores that are the furthest have the highest average sales and number of customers. Drop in Sales observed as the competition opens. We can clearly observe that most of the stores have their competition within low range.

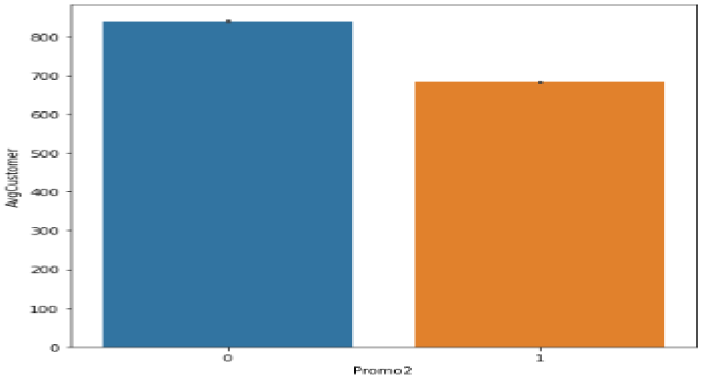
Effect of Promotion on sales & Customers





Promo2 & Promo2 Vs average sales and customers



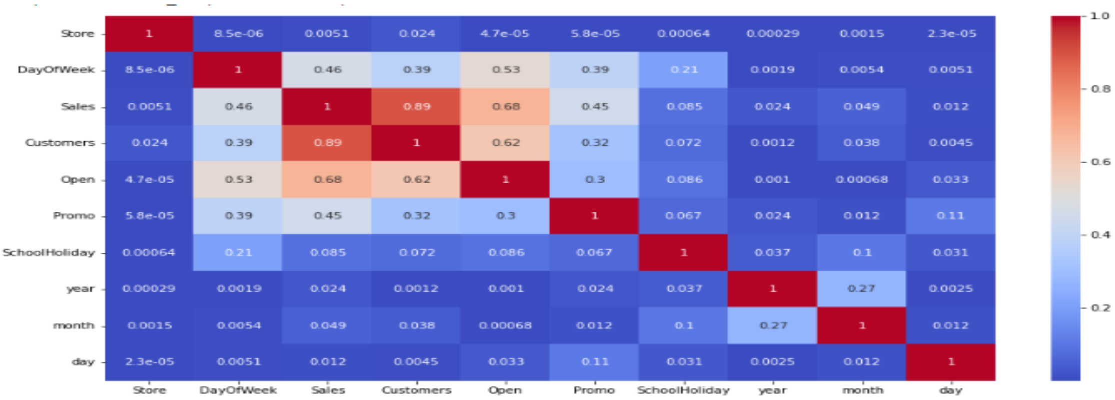


Customer over promotion we understand that initially those stores suffer from low sales and those continuous promotions show a tremending increase in the buying power of customers.

We can see that both Sales and Customers increase by a significant amount during Promotions. This shows that Promotion has a positive effect for a store.

.

Correlation Heatmap



	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday	year	month	day
Store	1	8.5e-06	0.0051	0.024	4.7e-05	5.8e-05	0.00064	0.00029	0.0015	2.3e-05
DayOfWeek	8.5e-06	1	0.46	0.39	0.53	0.39	0.21	0.0019	0.0054	0.0051
Sales	0.0051	0.46	1	0.89	0.68	0.45	0.085	0.024	0.049	0.012
Customers	0.024	0.39	0.89	1	0.62	0.32	0.072	0.0012	0.038	0.0045
Open	4.7e-05	0.53	0.68	0.62	1	0.3	0.086	0.001	0.00068	0.033
Promo	5.8e-05	0.39	0.45	0.32	0.3	1	0.067	0.024	0.012	0.11
SchoolHoliday	0.00064	0.21	0.085	0.072	0.086	0.067	1	0.037	0.1	0.031
year	0.00029	0.0019	0.024	0.0012	0.001	0.024	0.037	1	0.27	0.0025
month	0.0015	0.0054	0.049	0.038	0.00068	0.012	0.1	0.27	1	0.012
day	2.3e-05	0.0051	0.012	0.0045	0.033	0.11	0.031	0.0025	0.012	1

1. Customers and sales are positively correlated between 0.89
2. Sales and Promo (more than 0.3) actually correlate positively
3. Sales correlates with Competition Distance(more than 0.1), in a positive manner
4. Promo also effects sales positively

Interpretation:

We can first see the Sales and Promo (more than 0.3) actually correlate positively, since running a promotion increases that number .

Sales correlates with Competition Distance(more than 0.1), in a positive manner, like we said up the higher the competition distance the more sales per customer we do, which makes sense , the further our competition, the more monopolization Rossman can achieve in the region.

Additionally, the effect of promo to Sales like we said above as well(about 0.4), it did provoke a change in the buying pattern and increased it when continuous promotions were applied.

Feature Engineering

Since we need numerical variables for both our correlation Analysis and to feed the models, we need to transform what is not numerical to a numerical representation while keeping the logic behind it present. For this we did the below mentioned steps for model training:

1 Remove features with high percentages of missing values

2 Drop Subsets Of Data Which Might Cause Bias

3 Create a new variable "AvgSales" . Create a variable that calculates monthly average sales for each store.

4 Transform Variable "StateHoliday"

Conclusion of Exploratory Analysis:

At this stage, we got a solid understanding of the distributions, the statistical properties and the relationships of our variables. The next step is to identify what variables to model for training and to work on the modeling part of the project

Store Sales Prediction

Sales prediction is rather a regression problem than a time series problem. Practice shows that the use of regression approaches can often give us better results compared to time series methods. Machine-learning algorithms make it possible to find patterns in the time series. We can find complicated patterns in the sales dynamics, using supervised machine-learning methods.
Some of the most popular are tree-based machine-learning algorithms , e.g.,- Random Forest, Decision Tree etc . One of the main assumptions of regression methods is that the patterns in the past data will be repeated in future. In the sales data, we can observe several types of

Train-Test Split

We have split our variables into training and testing sets. We have performed this by importing train_test_split from the sklearn.model_selection library. It is usually a good practice to keep 70% of the data in your train dataset and the rest 30% in your test dataset.

Machine Learning Data Modeling (for our Prediction)

We need to build a Machine Learning model that will forecast future sales. Various methods of sales forecasting model that we will use in our project includes:

1. Linear Regression (OLS)

Ordinary Least Squares is a method which helps us estimate the unknown parameters in the Linear regression model. How does it estimate the parameters though? Well, it estimates the parameters by minimizing the sum of squared residuals. The way it does this is , it draws a line through the data points such that the squared differences between the observed values and the corresponding fitted value is minimized.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. ... A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable.

2. Lasso Regression

Least Angle regression(LARS). Basically, LARS makes leaps in the most optimally calculated direction without overfitting the model.

Algorithm:

- Normalize all values to have zero mean and unit variance.
- Find a variable that is most highly correlated to the residual. Move the regression line in this direction until we reach another variable that has the same or higher correlation.
- When we have two variables that have the same correlation, move the regression line at an angle that is in between (i.e., least angle between the two variables).
- Continue this until all of our data is exhausted or until you think the model is big and ‘general’ enough.

Mathematically, LARS works as follows:

- All coefficients, ‘B’ are set to 0.
- The predictor, xj is found to be most correlated to y.
- Increase the coefficient Bj in the direction that is most correlated with y and stop when you find some other predictor xk that has equal or higher correlation than xj.
- Extend (Bj, Bk) in a direction that is **equiangular** (has the same angle) to both xj and xk.
- Continue and repeat until all predictors are in the model.

3. Decision Tree Regression

Decision trees build regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

.

Model Comparison & Selection

There are two popular metrics used in measuring the performance of regression (continuous variable) models i.e MAE & RMSE.

- Mean Absolute Error (MAE): It is the average of the absolute difference between the predicted values and observed values.
- Root Mean Square Error (RMSE): It is the square root of the average of squared differences between the predicted values and observed values.

MAE is easier to understand and interpret but RMSE works well in situations where large errors are undesirable. This is because the errors are squared before they are averaged, thus penalizing large errors.
So, we'll choose RMSE as a metric to measure the performance of our models.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i}$$

Ai = actual value
Fi = forecast value
n = total number of observations

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

N = number of data points
yi = i-th measurement - Predicted sales.
y ̂(i) = the corresponding prediction - Sales.

Experiments and Metrics

Model	Training MAPE	Testing MAPE	Training RMSE	Testing RMSE	Model Score
Linear Regression	**17.016**	**17.214**	**1562.101**	**1573.522**	**0.7523**
LARS Lasso Regression	**17.017**	**17.213**	**1562.165**	**1573.894**	**0.7523**
Decision Tree Regression	**12.495**	**14.756**	**1216.922**	**1417.247**	**0.8497**

Model selection:

- As is shown in the result, among all models, **decision tree** works the best with the higher model score and least RMSE we have, and provides a reliable prediction of the sales.
- Linear regression and LARS Lasso Regression all have their own strengths and limitations.

.

Business Insights & Recommendations:

- Rossmann should focus on increasing the promotional offers per quarter for 1,2,3 and can minimize for 2.
- The most selling and crowded store type is 2
- Sales is highly correlated to the number of Customers.
- For all stores, Promotion leads to increase in Sales and Customers both.
- The stores which are opened during the School Holiday have more sales than normal days.
- ·More stores are opened during School holidays than State holidays.
- Rossman should try to focus on reducing the Promo offers for store type b during StateHolidays as there is no substantial increase in Sales.
- that people buy more beauty products during a Christmas celebration.
- Rossmann can divert some of the Promos from being offered on SchoolHolidays to No SchoolHolidays to maximise the Sales revenue.
- ·Absence of values in features CompetitionOpenSinceYear/Month doesn’t indicate the absence of competition as CompetitionDistance values are not null where the other two values are null.
- ·After analysing sales using Fourier decomposition, we found that there’s a little seasonality

component in the Sales data.

CONCLUSION AND FUTURE SCOPE :

That's it!

We reached the end of our exercise.

Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building

And all models have 74 to 85 accuracy levels its very good because its large data we getting 74 above