

Capstone Project – 2 Supervised ML - Regression Rossmann Sales Prediction

Team Alma Phoenix
Azhar Ali
Mohd Taufique
Pushpam Raghuvanshi

Al

Content

- 1. Problem Statement
- 2. Data Summary
- 3. Exploratory Data Analysis
- 4. Feature Engineering
- 5. Machine Learning Model Building
- 6. Challenges Faced
- 7. Conclusions
- 8. Recommendations

Problem Statement



- Rossmann operates over 3000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for upto six weeks in advance.
- Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.
- We are provided with historical sales data for 1,115 Rossmann stores.
- The task is to build Machine Learning model to forecast the "Sales" column for the test set.



DATA SUMMARY

Details of Datasets Provided



S. no.	Dataset	Variables	No. of Variables	No. of observations
1	Rossmann Stores Data - gives historical data including Sales	store, day of week, date, sales,customers, open, promo, state holiday, school holiday	9	1017209
2	store - supplement information about the stores	store, storetype, assortment, competition distance, competition open since month, promo2, promo2since week, promo2since year, promo interval	10	1115

Al

Important Variables

- 1. Sales The turnover for any given day (this is what we are predicting).
- 2. Open An indicator for whether the store was open or closed. 0 = closed, 1 = open.
- 3. Store type Differentiates between 4 different store models (a, b, c & d).
- 4. Assortment Describes an assortment level: a = basic, b = extra, c = extended.
- 5. Promo Indicates whether a store is running a promo on that day
- 6. Promo2 Promo2 is a continuing and consecutive promotion for some stores.



Important Variables(cont.)

- 7. Store A unique Id for each store.
- 8. Customer The number of customers on a given day.
- 9. Competition Distance Distance in meters to the nearest competitor store.
- 10. Promo Interval Describes the consecutive intervals Promo2 is started, naming the months the promotion is started a new.
- 11. Promo2Since [Year/week] Describes the year and calendar week when the store started participating in Promo2.

Missing Values



- Number of Missing values present in both the Datasets provided are:

0
0
0
0
0
0
0
0

Store	0
StoreType	0
Assortment	0
CompetitionDistance	3
CompetitionOpenSinceMonth	354
CompetitionOpenSinceYear	354
Promo2	0
Promo2SinceWeek	544
Promo2SinceYear	544
PromoInterval	544

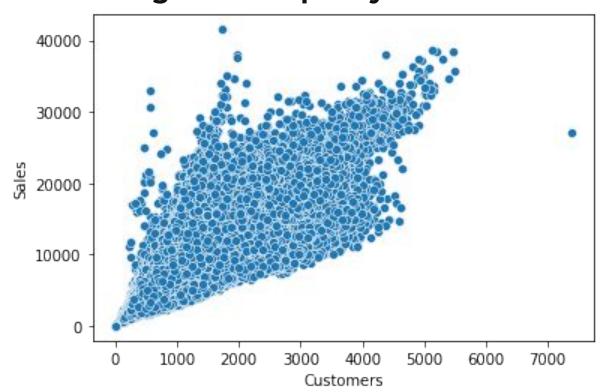


EXPLORATORY DATA ANALYSIS

Scatterplot - Sales and Customer



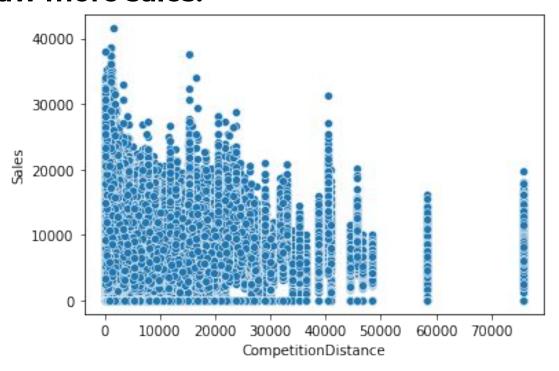
- Scatterplot shows that relation between Sales and customer is sort of linear. Sales is increasing with the number of Customers increasing which is pretty obvious.





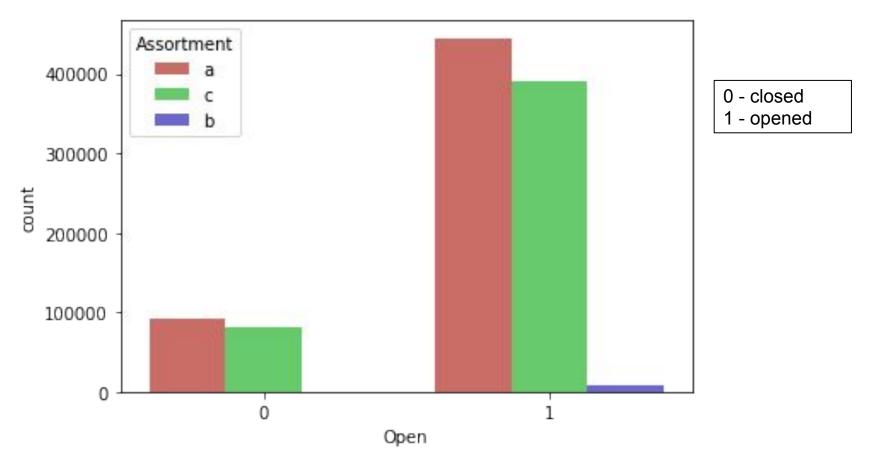
Scatterplot - Sales and Competition Distance

- 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.



Countplot - Store Open/Closed

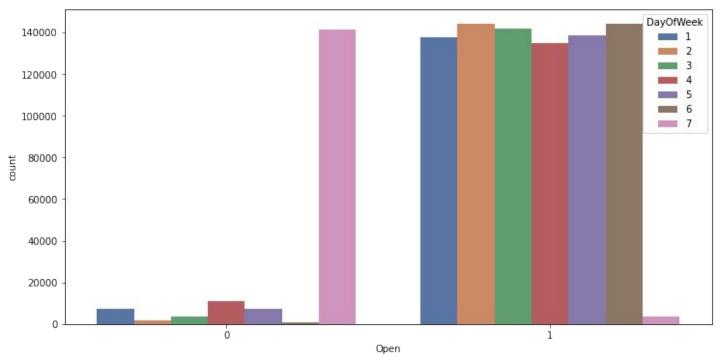






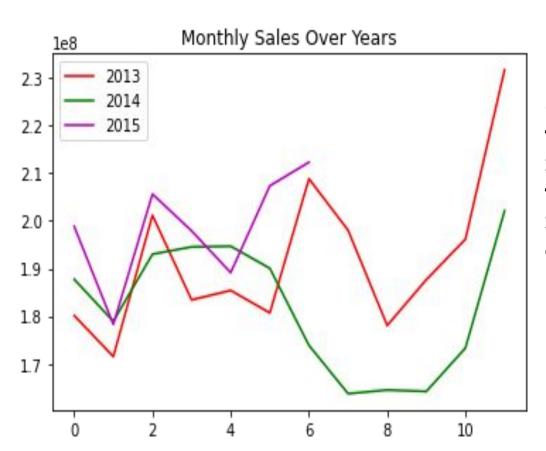
Store Open/Closed over Day of Week

- This countplot clearly shows that majority of stores are closed on sunday. Some stores were also closed on other days of the week may be due to public holidays & refurbishment, as stores are usually closed on public holidays and are open during school vacations.



Average Sales Over Year/Month

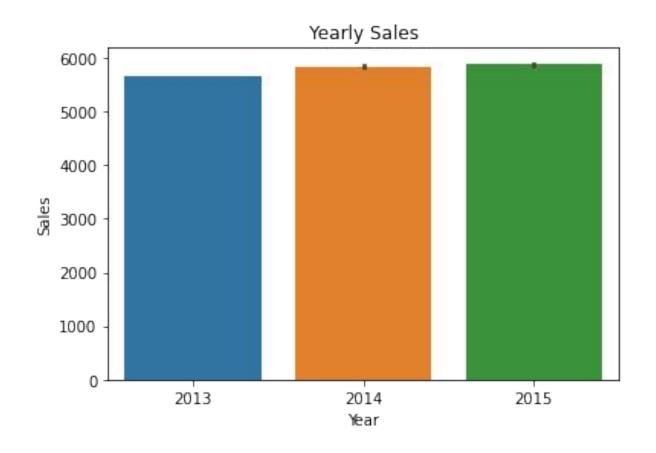




- Sales rise up by the end of the year before the holidays. Sales for 2014 went down there for a couple months - July to September, indicating stores closed due to refurbishment.

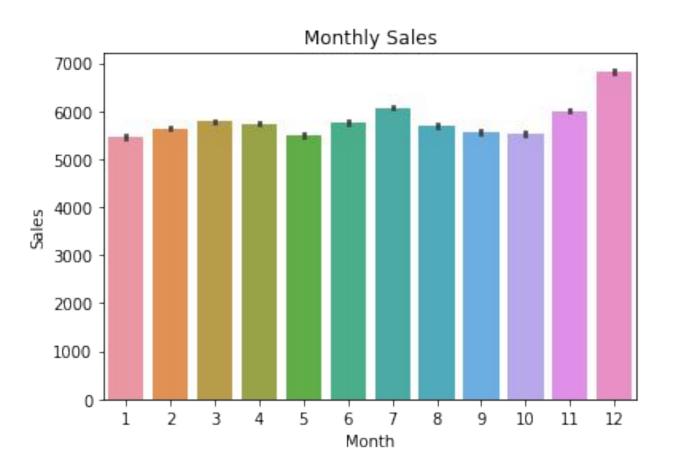
Total Sales in successive Years





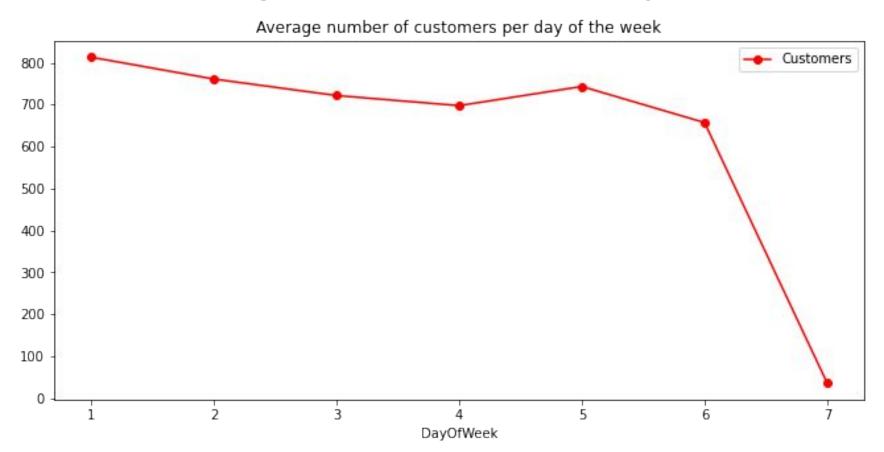
Total Sales in Months





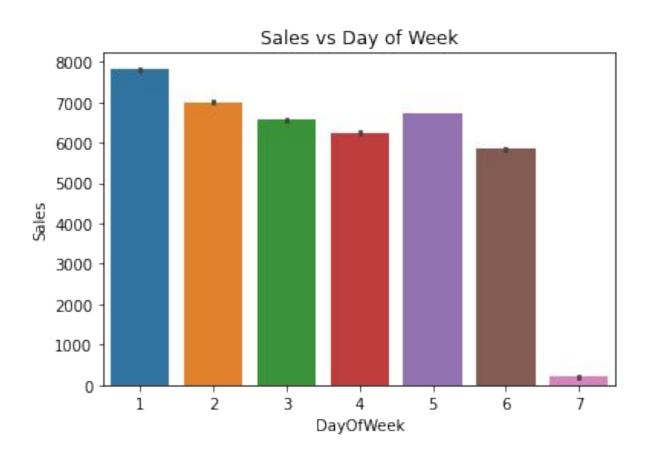


Trend - Average Customer per Day of Week



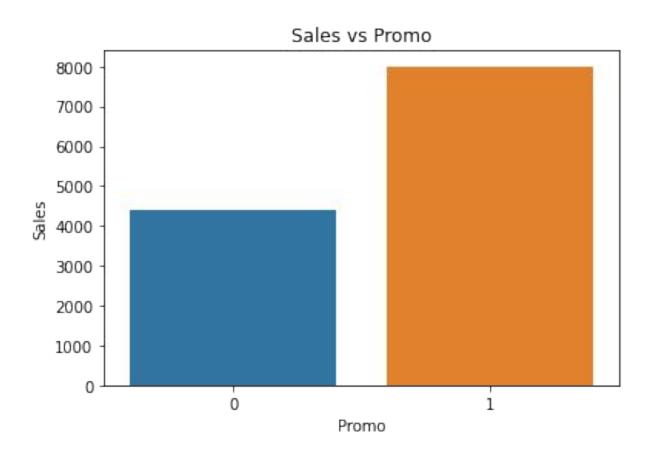
Total Sales on Weekdays





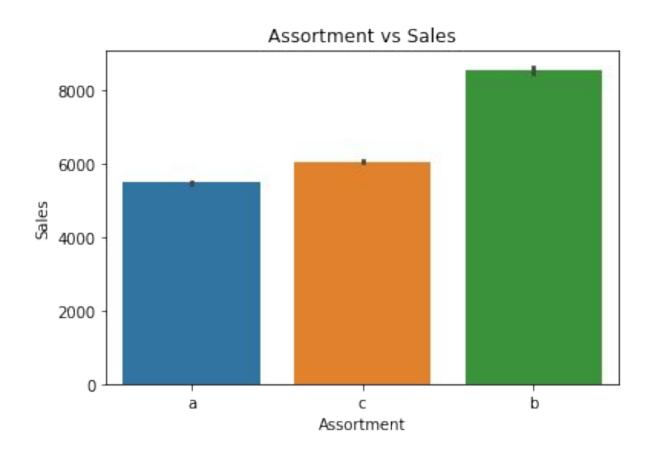
Sales VS Promo





Sales vs Assortment

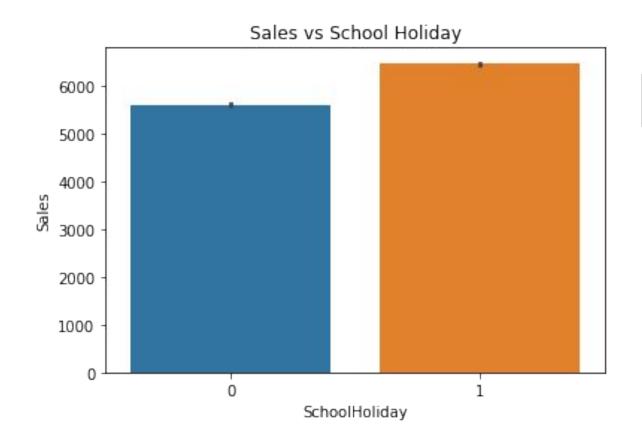




- a basic
- b extra
- c extended

Sales vs School Holiday

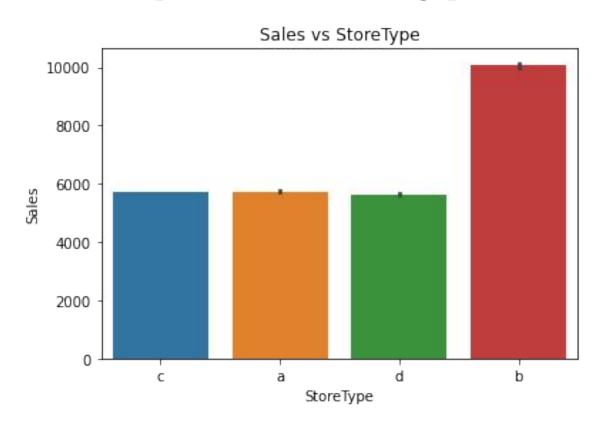


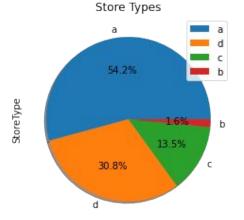


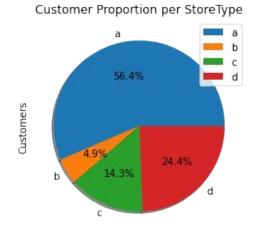
- 0 Not Holiday
- 1 School Holidays

Proportion of Store Types and Total A

Sales per Store Type









FEATURE ENGINEERING



Replacing Missing Values

- There are many Missing values in our Store Dataset.

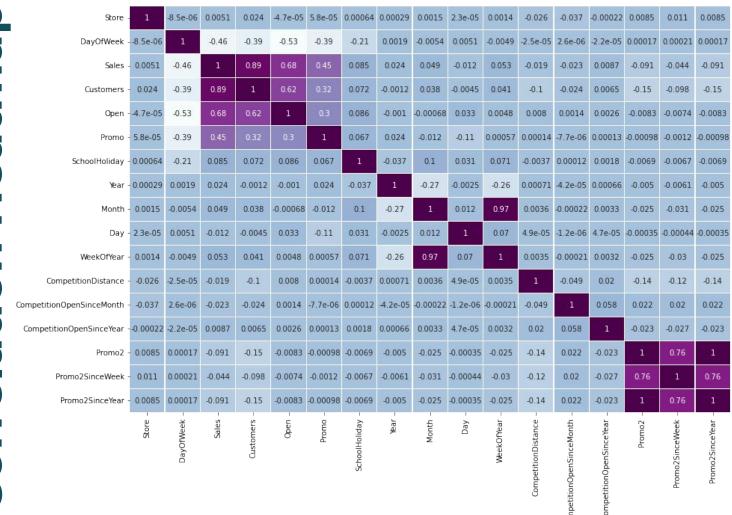
Feature Name	Number of Missing values	Replaced with	Remark	
Promo2SinceWeek	544		Not much	
Promo2SinceYear	544	o	Information given about	
Promointerval	544		the features.	
CompetitionOpenSinceMonth	354	Mode		
CompetitionOpenSinceYear	354	Mode		
CompetitionDistance	3	Median	As they are only 3.	



Changing Datatype of Columns

- In Store Dataset StoreType, Assortment and PromoInterval is of Object type.
- StoreType and Assortment have values [a,b,c,d] and [a,b,c] respectively.
- Changing these to Numerical values [0,1,2,3] and [0,1,2] respectively.
- PromoInterval has values [0, 'Feb,May,Aug,Nov', 'Jan,Apr,Jul,Oct', 'Mar,Jun,Sept,Dec']
- Changing these to dummy variables.

T



- 0.8

- 0.6

- 0.4

0.2

- 0.0

- -0.2

- -0.4

Removing Multicollinearity

Al

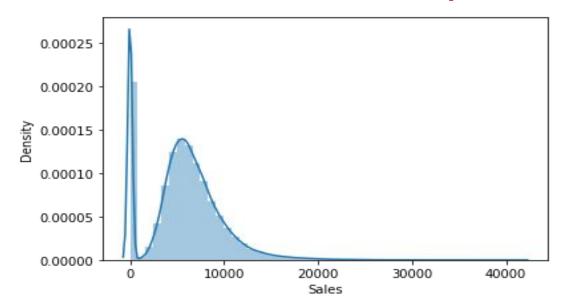
- Removing the features which is having VIF>10 because it will affect & interpret the result.
- VIF <= 10 is usually preferred as this can easily explain the variance of 90% i.e, R-square becomes 90%.(VIF=1/1-R^2)

	variables	VIF	
0	Store	3.627575	
1	DayOfWeek	4.513547	
2	Customers	4.339790	
3	Promo	1.946273	
4	StateHoliday	1.003985	
5	SchoolHoliday	1.247951	
6	Day	3.847661	
7	StoreType	1.916142	
8	Assortment	2.049916	
9	CompetitionDistance	1.532510	
10	Competition Open Since Month	6.574554	
11	Promo2	4.752803	
12	Promo2SinceWeek	3.737566	

Filtering Rows(Records)



- Filtering records where stores are closed as they won't generate any Sales.
- Filtering records where stores has Sales equal to 0.





MACHINE LEARNING MODEL BUILDING

Evaluation Metrics

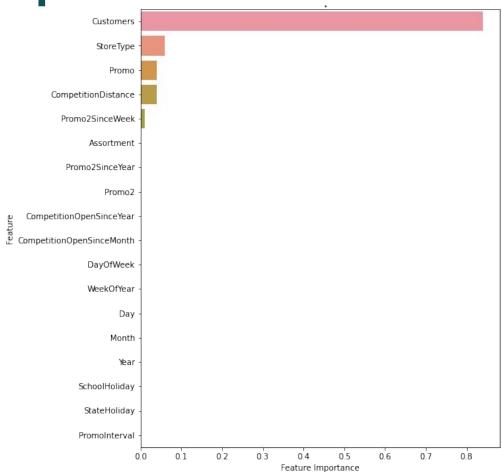


ML Model predict sales for stores which are open and when there is some sales because there is no sales when store is closed.

	Linear Regression (OLS)	Lasso Regression (L1)	Ridge Regression (L2)	Decision Tree	Random Forest	Random Forest with hyperparameter tuning & CV
RMSE	1519.132	1520.620	1519.849	1433.262	1126.660	<u>513.893</u>
MAPE	15.902	15.920	15.908	15.478	12.442	4.993
R2	0.761	0.7607	0.7609	0.7609	0.8686	0.9727

Feature Importance





Actual vs Predict Forest Hyperpara Tuned Model

168863

168864

168865

168866

168867

11843

11961

4657

9680

4252

2581

3757

3540

168868 rows x 2 columns

		ues f	 			
	actual	predicted				
0	6792	6277.666667				
1	11585	11038 933333				

11386.666667

10756.933333

4456.000000

9094.066667

4434.333333

2466.466667

3371.000000

3652.933333



Model Selection

- By Looking at the evaluation metrics obtained on implementing different sort of regression model, we decided to go with the Random Forest Tuned model. The maximum R^2 was seen in tuned Random Forest model with the value 0.97267. It means our best accurate model is able to explain approx/almost 97% of variances in the datasets.
- Based on our model; Customer, store Type, Promo & Competition Distance are the most impactful features which are driving the sales more as compared to other features present in the dataset.

Challenges Faced

Al

- Understanding the meaning of some columns.
- Handling Large amount of Sales Data.
- Dealing with Null values, as there are many Null values.
- Understanding the business model of Retail Sales that how they work.
- Also, forming different graphs to show insights from the dataset and to summarize the information and communicate the results and trends to the reader successfully.
- Dealing with Categorical columns to make them numerical for make use in ML model building.
- Selecting appropriate Model to fulfill the purpose.



Conclusions

- From the sales and customer scatterplot, the relationship is sort of linear ie sales is increasing with number of customers increasing which is obvious.
- Stores with Assortment level 'b' has the highest sales.
- Approx. 50% stores are of type 'a'. There are very few stores of type 'b'.
- Store type 'b' has the highest sales and all other store types 'a','c','d' has nearly equal sales.
- December records the highest monthly sales. This may be due to Christmas and New Year.
- Sales is more when promos/offers are running on stores.



Recommendations from our Analysis

- More stores should be encouraged for promos.
- Store type 'b' should be increased in number.
- There is seasonality involved. Hence, the stores should be encouraged to promote and take advantages of the holidays.



Thank You

Hope our best model will help the Store managers to predict the daily sales accurately in advance.

CHEERS ₩ !!!

Team Alma Phoenix