Capstone Project - 2
Regression Analysis
Rossman Sales Prediction

Presented by:
Ashutosh Sharma
Kumar Aman

Content

- What is Rossman?
- Understanding The data
- Eda
- Data Cleaning and Preprocessing
- Model building
- Model Evaluation
- Conclusion

What is Rossman?

- Dirk Rossmann GmbH, commonly referred to as Rossmann, is one of the largest drug store chains in Europe with around 56,200 employees and more than 4000 stores.
- The company was founded in 1972 by Dirk Rossmann with its headquarters in Burgwedel near Hanover in Germany.
- In 2019 Rossmann had more than €10 billion turnover in Germany, Poland, Hungary, the Czech Republic, Turkey, Albania, Kosovo and Spain.
- The Rossmann family owns 60% of the company. The Hong Kong-based A.S. Watson Group owns 40%, which was taken over from the Dutch Kruidvat in 2004.
- The product range includes up to 21,700 items and can vary depending on the size of the shop and the location.
- In addition to drugstore goods with a focus on skin, hair, body, baby and health.
- Rossmann also offers promotional items ("World of Ideas"), pet food, a photo service and a wide range of natural foods and wines.

src- wikipedia

Data Summary

Data Includes two major datasets in CSV format

1. Rossman Stores Data.csv -historical data including Sales
2. Store.csv -supplemental information about the stores

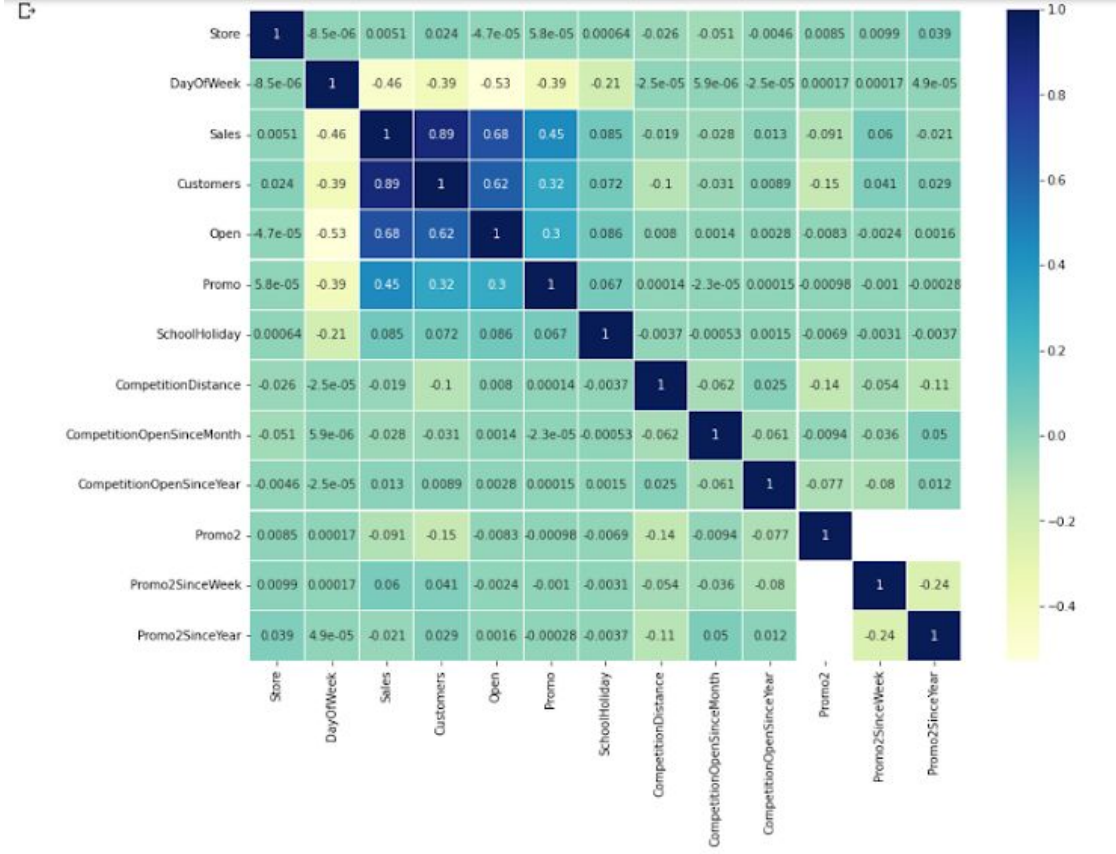
Type	Feature	Dataset	comment
Categorical	Storetype	Stores.csv	a, b, c, d
	stateholiday	Rossmann Stores Data.csv	a = public holiday, b = Easter holiday, c = Christmas
	assortment	Rossmann Stores Data.csv	a = basic, b = extra, c = extended
	PromoInterval	Stores.csv	
Unique	Store	Stores.csv	Unique ID for each Store
	Date	Rossmann Stores Data.csv	
Numeric-continous	**Sales**	**Rossmann Stores Data.csv**	Target
	customers	Rossmann Stores Data.csv	Number of customers on a partiular day
	competetiondistance	Stores.csv	Distance to a nearest competitor
Numeric-Categorical	Day of Week	Rossmann Stores Data.csv	Ranges 1-7
	CompetetionOpenSinceMonth	Stores.csv	
	CompetetionOpenSinceYear	Stores.csv	
	Promo2SinceWeek	Stores.csv	week from which promotion was initiated
	Promo2SinceYear	Stores.csv	year from which promotion was initiated
Binary	Open	Rossmann Stores Data.csv	0: Closed, 1: open
	Promo	Rossmann Stores Data.csv	
	Promo2	Stores.csv	
	SchoolHoliday		

Data Understanding?

- A total of 1017209 rows and 8 features were present in Rossman stores data.csv and
- A total of 1115 rows and 9 features were present in stores.csv
- Store feature was common among both dataset with unique store id, after merge a total of 17 columns were left.
- A total of 3 categorical variables were present in PromoInterval, with starting months.
- Few columns like Promo2sinceweek, Promo2sinceyear, PromoInterval have around 49% null values.
- CompetetionopenSinceMonth and CompetetionOpenSinceYear have around 31% null values and CompetetionDistance have 0.26% null values.
- Outliers were present in Target variable as well as other numerical variables too including CompetetionDistance, CompetitionOpenSinceMonth/Year, Promo2sinceweek/year etc .
- For few features a measure of central tendency imputation were done for few null imputation was performed.

Exploratory Data Analysis

Corelation Heatmap



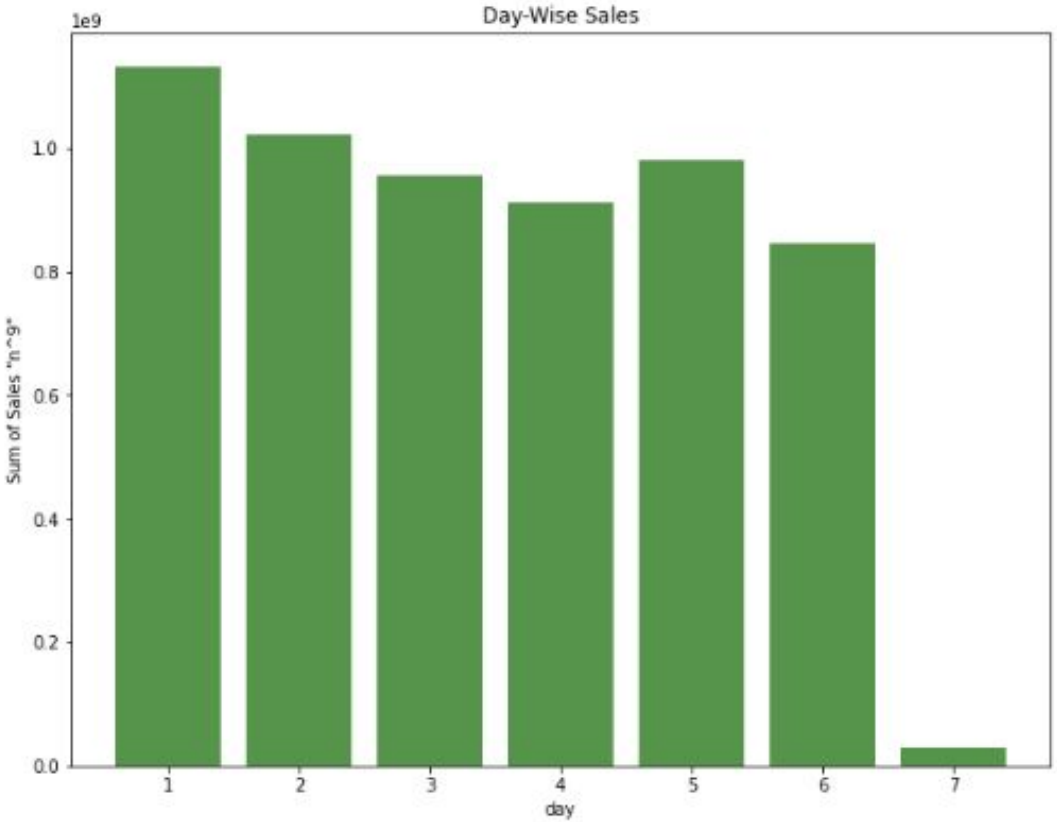
Key-takeaways:

- Columns like Customer, Open, Promo Shows nice correlation with Target Variable.
- Features like DayOfWeek, Promo2 does not have significant effect on Sales feature.
- Multicollinearity can be seen in Promo2, Promo2SinceWeek, Promo2SinceYear

Que1: On which day the cumulative sales were highest?

Key-takeaways:

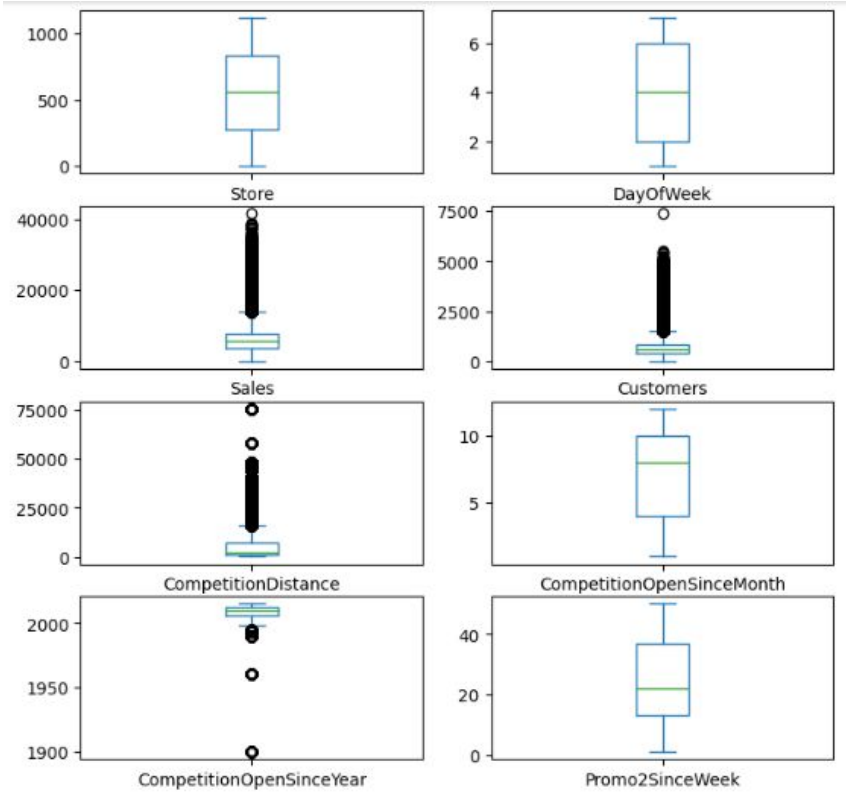
- On Day 1 Cumulative sales are very high on Day1 and keeps on decreasing till day 4.
- Day 5 Shows hike.
- Day 7 Sales are very Low Compared to other 6 days
- There might be Weekend holiday of the store on 7th day in most countries.



Que2: Comparing Data distribution of various columns

Key-takeaways:

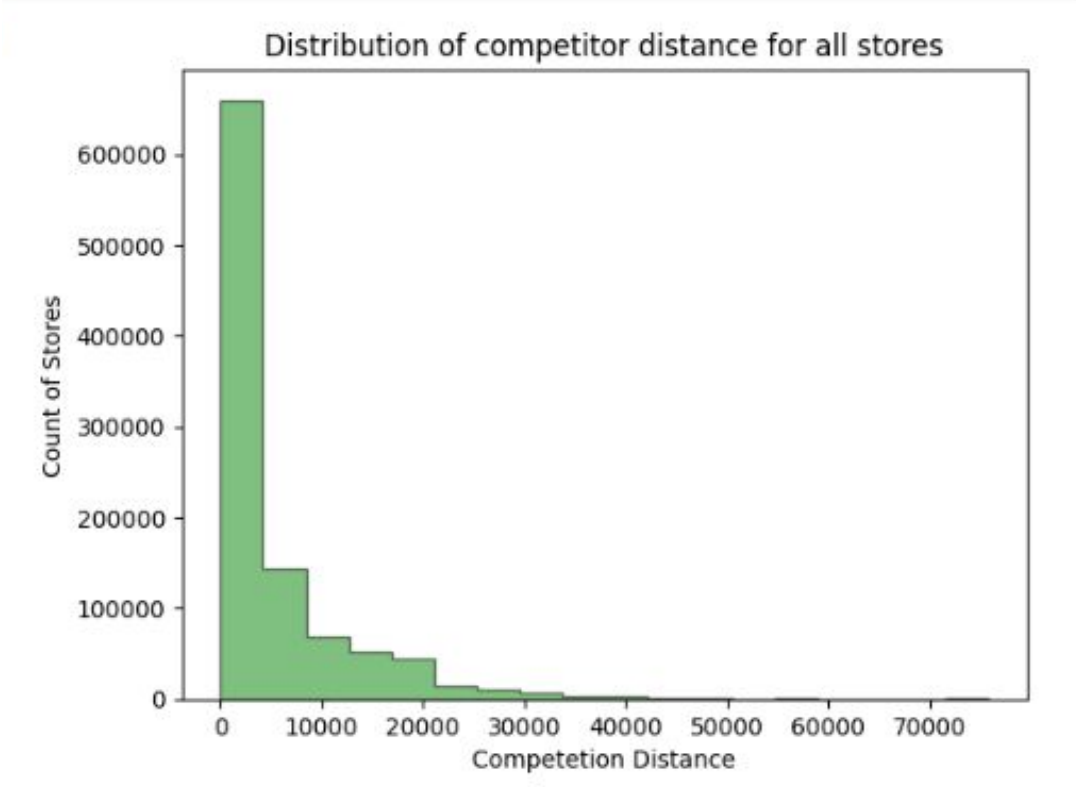
- Sales Feature shows high number of outliers
- CompetetionOpenSinceMonth shows high number of value in 8th month (August).
- CompetetionOpenSinceYear shows few values below 2000 but some very robust outliers are present as near 1900 and 1950
- Features as Customer and Competition Distance also have Outliers but other features shows good distribution.



Que3: Distribution of competitor distance for all stores

Key-takeaways:

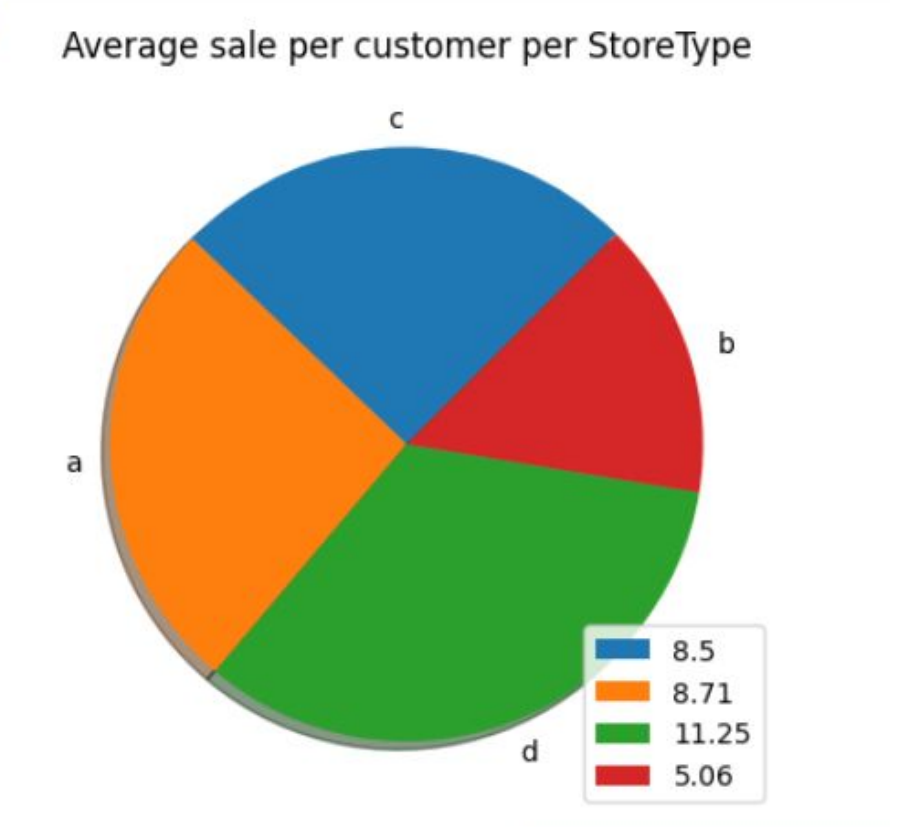
- Competetion Distance Shows a Pareto Distribution
- Competetion Distance among Stores lies between 0-70000 meters
- Mostly Competetion Distance Lies between 0-10000



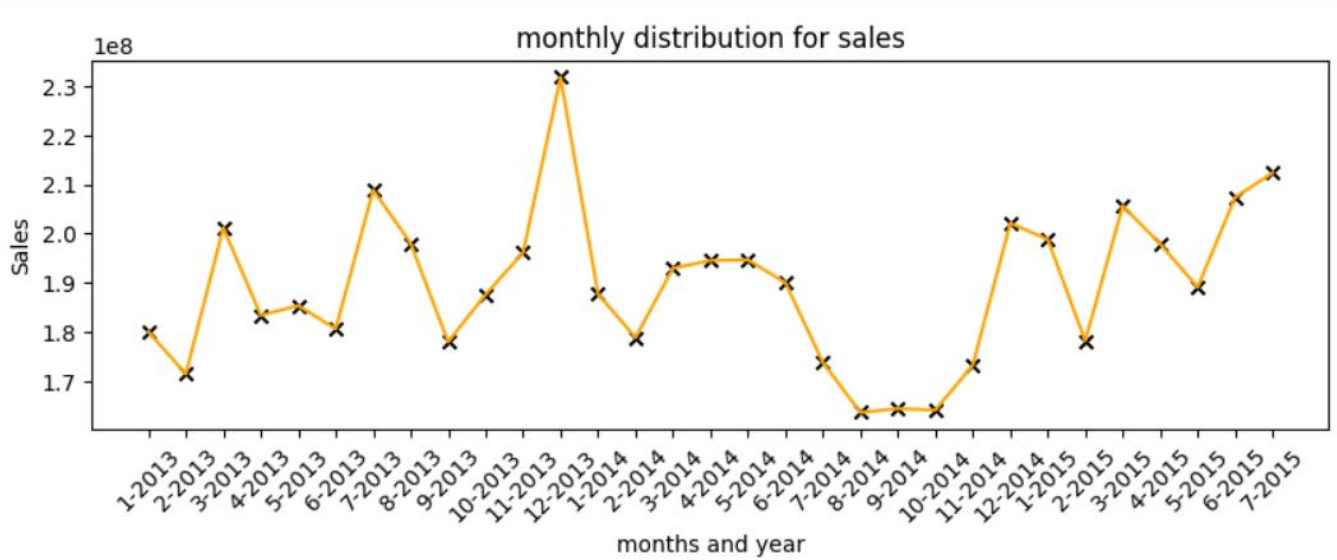
Que4: What is the average sale per customer per StoreType?

Key-takeaways:

- Highest Per Customer Sale is provided by Store D.
- Lowest Per customer sale is 5.06 through Store B.
- Store A and C shows similar sales per customer
- Company should Focus more on Store B as it shows low per customer sale value.



Que5: Monthly distribution for sales



Key-takeaways: Non Stationary Data can be seen for Monthly Sales distribution.
November 2013 was showing highest number of Sale overall.
Monthly Sales of the Drugs may depend upon multiple other factors as it shows high variance.

Que6: Anova analysis for sales over assortment

A one-way ANOVA has the below given null and alternative hypotheses:
H0 (null hypothesis): μ1 = μ2 = μ3; The means of all the Sales of all assortment are equal
H1 (null hypothesis): There will be at least one assortment type whose mean Sales differs from the rest

```
F_onewayResult(statistic=9256.967502400194, pvalue=0.0)
```

Key-takeaways: The F statistic and p-value turn out to be equal to 9256.96 and 0.0 respectively. Since the p-value is less than 0.05 hence we would reject the null hypothesis. This implies that we have sufficient proof to say that there exists a difference in the performance among sales of various assortment types

Feature Engineering

- CompetitionOpenSinceYear, CompetitionOpenSinceMonth, Promo2SinceWeek and Promo2SinceYear were having around 50% null values so replaced those values by 0.
- Features like StateHoliday were having mixed datatypes thus they were made of a single dtype
- (avoided one hot encoding as it creates multicollinearity between variables)
- Performed various encoding techniques like(binary and label encoding) On Others features like (StoreType, Assortment) similar to StateHoliday.
- Mean Imputation on CompetitionDistance as Zero imputation would not be a better option here.
- Converted data to numeric data type.

Feature Engineering

- Applied VIF:

	feature	VIF
0	Store	1.007318
1	DayOfWeek	1.721571
2	Date	514.886598
3	Sales	9.786515
4	Customers	9.664344
5	Open	2.503731
6	Promo	1.443125
7	SchoolHoliday	1.095271
8	Promo2	2.431292
9	Promo2SinceWeek	2.384532
10	public holiday	1.926120
11	Easter holiday	4.384840
12	Christmas holiday	4.808236
13	StoreType_0	2.468508
14	StoreType_1	1.526559
15	StoreType_2	1.553359
16	Assortment_0	67.134926
17	Assortment_1	68.191673

- Date was showing high Multicollinearity so removed that.
- Removed outliers of the Target variable with sales lower than Q1-1.5*IQR and sales above Q3+1.5*IQR

Data Preprocessing

Steps

1. Defined X and y as Independent and Target Variable
2. Scaled the independent variables for better accuracy Using MinMax Scaler
3. Train Test Split with a Test size of 20%

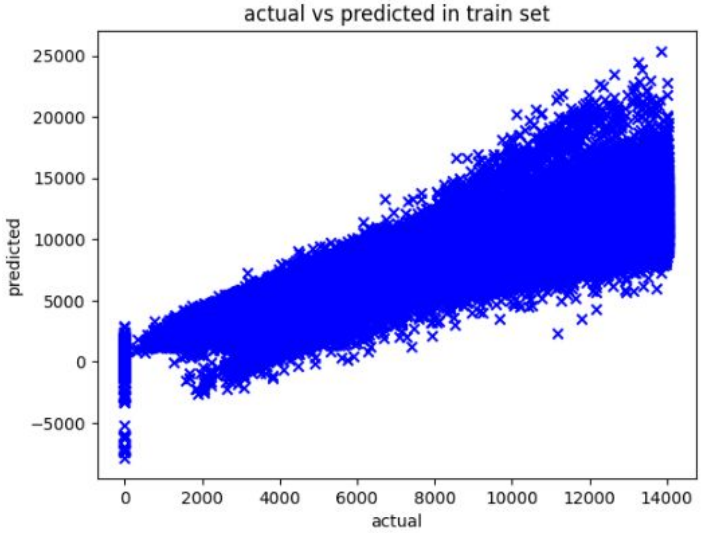
Model Building

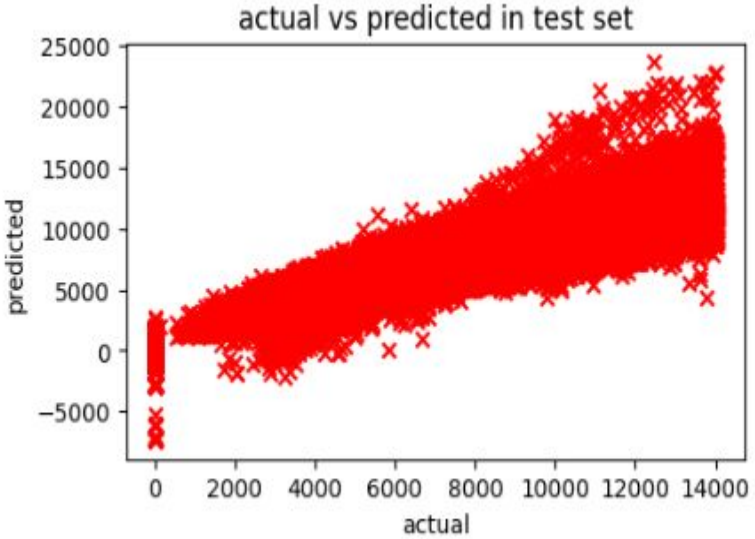
The Most Important Step again is to Build a Sustainable and trustworthy model

Lets Train Computer by Experience

- **OLS Regression**
- **OLS with k fold CV with 5 folds**
- **L1 Reg.. Lasso Regression**
- **L2 Reg.. Ridge Regression**
- **GBR (Gradient Boosting Regressor)**

Train vs Test Prediction in OLS





Model Evaluation

Evaluation metric / Model	MSE(test)	MAE(test)	Train Accuracy	Test Accuracy	Rank
OLS	1168946.14	812.52	0.8984	0.8967	2
OLS with Grid Search			0.8984		
Lasso (L1 Reg..) 5 fold CV	1173761.77	814.37	0.898	0.8963	4
Ridge (L2 Reg..) 5 fold CV	1170940.73	813.52	0.8982	0.8965	3
GBR (Gradient Boosting Regression)			0.9614	0.9607	1

Conclusion

- Some features were extremely important for the Prediction like Customer, Promo, Open etc
- OLS performed better than lasso and ridge which shows that there was negligible variance and bias in the model.
- Due to Large Dataset the model was trained well and reduced the occurrence of under-fitting and over-fitting.
- Out of total 4 models Gradient Boosting Regressor performed best as it retrain the model putting extra weightage on wrong predictors.
- On Sunday the market might gets close, resulting in Extremely less sales.
- The Data could also be used for Time Series Analytics but as the data is old of 2015, now it is very less likely for the firm to use this in future prediction.

Thank You