

Welcome To Our Machine Learning Regression

Capstone Project -- 2



Retail Sales Prediction

Presented by

Ashwani Kumar

Machine Learning Projects In Retail



Content

- 1. Problem Statement
- 2. Data Summary
- 3. Data Preprocessing
- 4. Exploratory Data Analysis
- 5. Feature Engineering
- 6. Model Implementation
- 7. Conclusion

Problem Statement

- Rossmann operates over 3000 drug stores in 7 European countries.
- 2. Provided with historical sales data for 1,115 Rossmann stores. The sales are influenced by many parameters and task is to forecast the "Sales" for 6 weeks in advance.



Data Summary

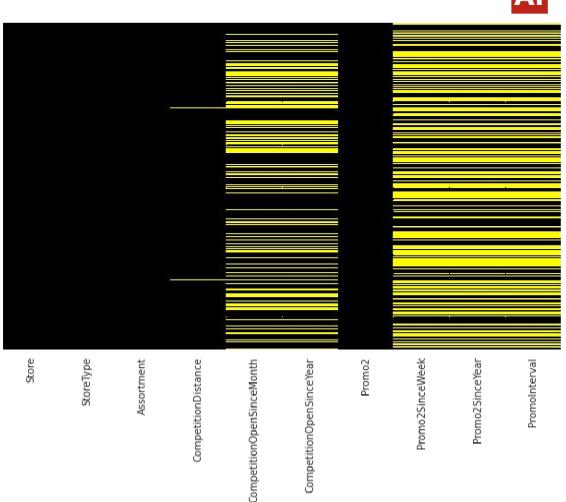
We have two datasets. Rossman store data is for years 2013, 2014 and 2015 with 10,17,209 observations on 9 variables. Stores data with 1115 observations on 10 variables. Some important features are:-

- 1. Customer: The number of customers on a given day in a store.
- 2. Date :- Showing dates for observations.
- 3. State Holiday: Indicating a state holiday.
- 4. Store Type: Differentiate between 4 different store models (a,b,c,d).
- 5. Assortment : Describes an assortment level i.e a : basic, b : extra and c : extended.
- 6. Competition Distance : Distance in meters to the nearest competition store.
- 7. Promo :- Indicates whether a store is running a promo on that day.

Columns having >30% null values are dropped.

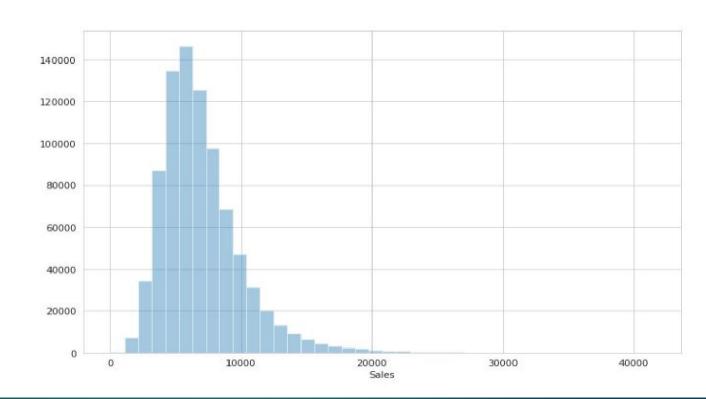
Null values in 'Competition Distance' are imputed with median of feature.

Removing those stores observations that are temporarily closed (~ 17.3K) & stores generating zero sales.

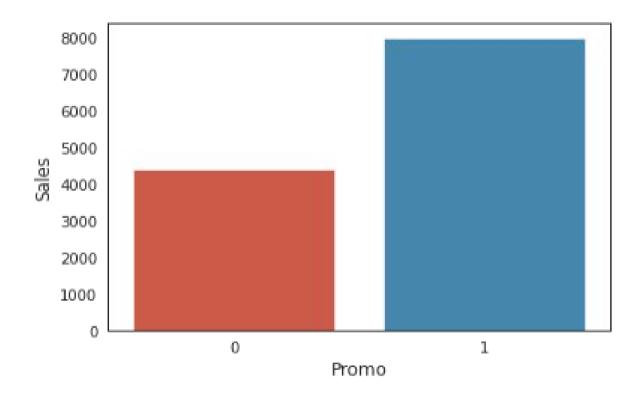


Exploratory Data Analysis

Sales are normally distributed with slightly right tail skewed.



Impact of Promo on sales: Sales Are nearly doubled High When Promo is Running



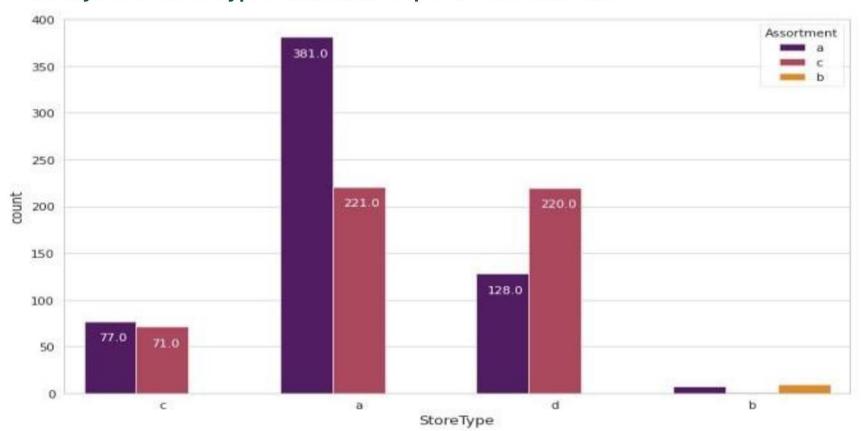
EDA (contd..)
Heatmap for merged dataset

		•											
Store	1	8.5e-06	0.0051	0.024	4.7e-05	5.8e-05	0.00064	0.00029	0.0015	2 3e-05	0.0014	0.026	0.0085
DayOfWeek	8.5e-06	1	0.46	0.39	0.53	0.39	0.21	0.0019	0.0054	0.0051	0.0049	2.5e-05	0.00017
Sales	0.0051	0.46	1	0.89	0.68	0.45	0.085	0.024	0.049	0.012	0.053	0.019	0.091
Customers	0.024	0.39	0.89	1	0.62	0.32	0.072	0.0012	0.038	0.0045	0.041	0.1	0.15
Open	4.7e-05	0.53	0.68	0.62	1	0.3	0.086	0.001	0.00068	0.033	0.0048	0.008	0.0083
Promo	5.8e-05	0.39	0.45	0.32	0.3	1	0.067	0.024	0.012	0.11	0.00057	0.00014	0.00098
SchoolHoliday	0.00064	0.21	0.085	0.072	0.086	0.067	1	0.037	0.1	0.031	0.071	0.0037	0.0069
Year	0.00029	0.0019	0.024	0.0012	0.001	0.024	0.037	1	0.27	0.0025	0.26	0.00071	0.005
Month	0.0015	0.0054	0.049	0.038	0.00068	0.012	0.1	0.27	1	0.012	0.97	0.0036	0.025
Day	2.3e-05	0.0051	0.012	0.0045	0.033	0.11	0.031	0.0025	0.012	1	0.07	4.9e-05	0.00035
WeekOfYear	0.0014	0.0049	0.053	0.041	0.0048	0.00057	0.071	0.26	0.97	0.07	1	0.0035	0.025
ompetitionDistance	0.026	2.5e-05	0.019	0.1	0.008	0.00014	0.0037	0.00071	0.0036	4.9e-05	0.0035	1	0.14
Promo2	0.0085	0.00017	0.091	0.15	0.0083	0.00098	0.0069	0.005	0.025	0.00035	0.025	0.14	1
	Store	DayOfWeek	Selection	Customers	Open	Pramo	SchoolHoliday	Mear	Month	Day	WeekOffear	ompetitionDistance	Promo2



ΑI

Analysis of Store Types with their respective assortment.

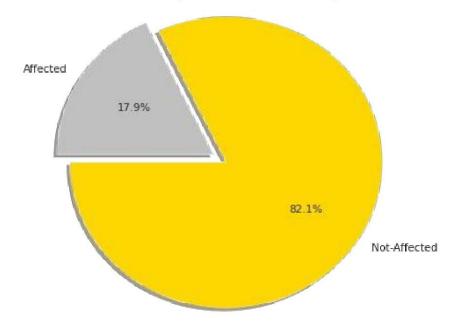


EDA (contd..)

School and State holidays effect on sales

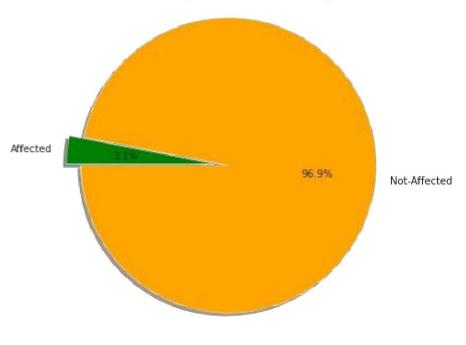
As we can see in the Piechart Sales affected by School Holiday is 18% and Mainly Sales aren't afffected by School Holiday

Sales Affected by Schoolholiday or Not?



As we can see in the Piechart Sales affected by State Holiday is only 3% means Sales aren't afffected by State Holiday

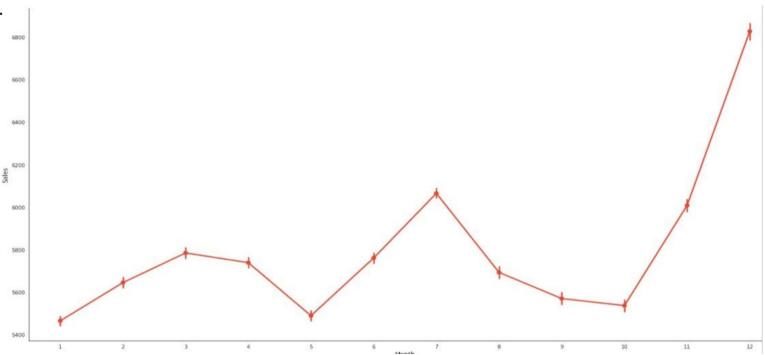
Sales Affected by State holiday or Not?



ΑI

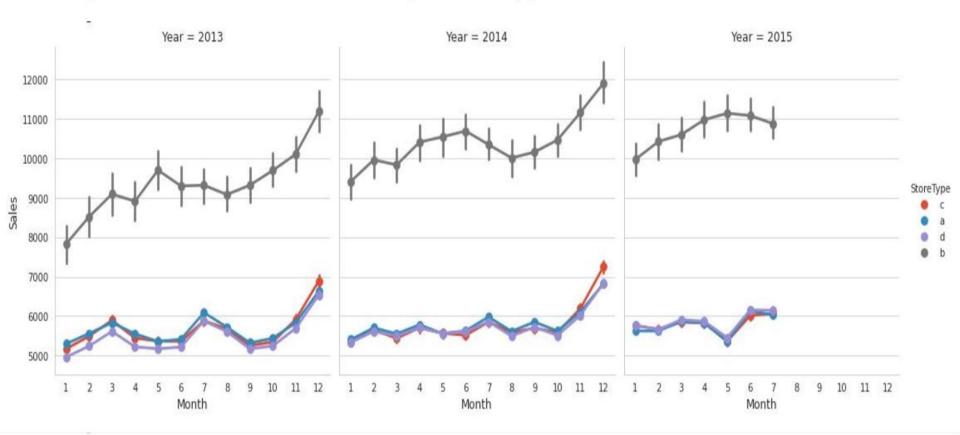
Monthly trends in Sales:

As We can see that In the month of November and Specially in December Sales is increasing Rapidly every year on the christmas eve.



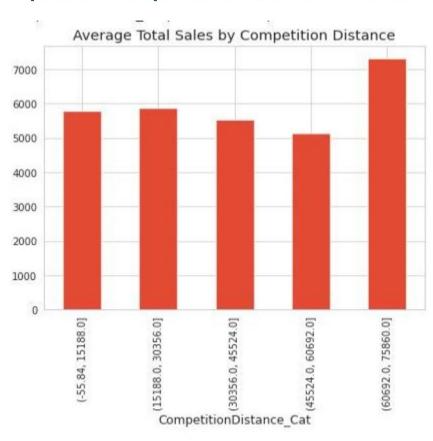
Al

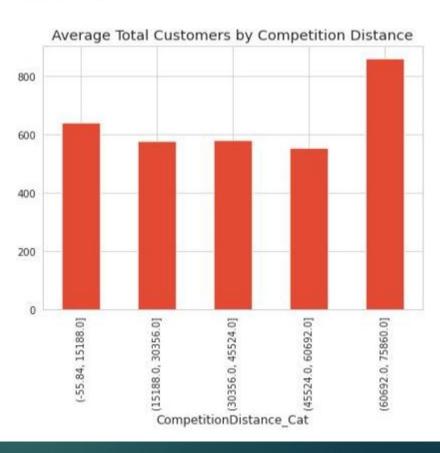
Yearly Distribution of Sales according to store types



Al

Impact of Competition Distance on Sales and Customers





EDA (summary)

- 1. Sales are highly correlated to customers.
- 2. Stores opened on 'State Holiday' makes a good amount of sales.
- 3. There is no such significant difference in sales on 'School Holidays'.
- 4. Even though store type 'b' has very less number of stores but these are outperforming other store types in terms of sales and avg customers.
- 5. Sales are consistent for the second quarter of the year but it starts increasing in the last quarter.



Feature Engineering

- 1. Extracting week, month, year from Date and adding them in dataset.
- 2. Merging both dataset.
- 3. One hot encoding for Storetype, Assortment.
- 4. Splitting dataset into Training and Test set and applying MinMaxScaler for scaling dataset.



Models Implemented

- 1. Linear Regression (Baseline Model)
- 2. Lasso Regression
- 3. Decision Tree Regression
- 4. Decision Tree Regression (with hyperparameters)
- 5. K-Nearest Neighbors Regression
- 6. Random Forest Regressor

Model Evaluation

MODEL	TRAINING SCORE	TESTING SCORE
Linear Regression	0.780750	0.782392
Lasso Regression	0.780731	0.780769
Decision Tree Regression	0.99996	0.915942
Decision Tree Regression (with hyperparameters)	0.963506	0.935417
K-Nearest Neighbors Regression	0.73722	0.71665
Random Forest Regressor	0.993783	0.956520



Conclusion

Our model shows that Customers, Competition distance, Store type are some of the most important features in our sales prediction. We need to focus on these aspects to maximize our profits for the next 6 weeks.

THANK YOU

