Capstone Project

Retail Sales Prediction
By-Aditi Rajguru

Problem Statement

Review Sales Prediction is a data of Rossmann Stores which operates over 3000 drug stores over 7 European Countries. Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance.In the datasets there are various fields like promotions,competition,school and state holidays,etc which are the affecting factors for Store sales and The historical sales data for 1,115 Rossmann stores was provided.In this project we are working with process like problem definition,data wrangling,feature engineering,data visualization and selecting model using regression.

We have tried to solve each and every possible errors in our project. In this presentation we are presenting our journey step by step including conclusions and challenges we faced when we were working on this project.

Table of Contents

- Data Set Used & Feature Representation.
- Introduction.
- Feature Engineering.
- EDA with Data Visualizations.
- Feature Selection.
- Applying models.
- Model Observation and Validation.
- Challenges Faced.
- Conclusion.

Datasets Used

IN this analysis we have used the following datasets

1)Rossmann_Store_Data : Contains historical sales data like store,date,sales,customers,School and State holiday,etc.

2)Store Data: Contains supplemental information about the stores like Assortment,Promo2,Competition Distance,etc.

3)Merged both with Left Joint : This dataset is actually formed by merging the above two dataset in order to map information from one dataset to another.

4) Final merged dataset: This dataset is subset of merged datasets with only specific columns and dropping of columns which we don’t need.

Features Presentation:

Data give us the following features:

- 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.
- StateHoliday - 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.
- SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools.
- 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.
- CompetitionOpenSince[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.

Why is sales prediction important for Rossmann Store?

A sales forecast is an estimation of future sales. This estimation can be based on past values, economic indicators, seasonality, weather forecasts, promo, Assortment of product etc.with forecasting we can answer Questions like.

1. How much stock should be ordered?
2. How much revenue can be expected in upcoming Year?
3. High accuracy of sales with respect to unique circumstances and conditions.

Data PreProcessing.

Feature Transformation:

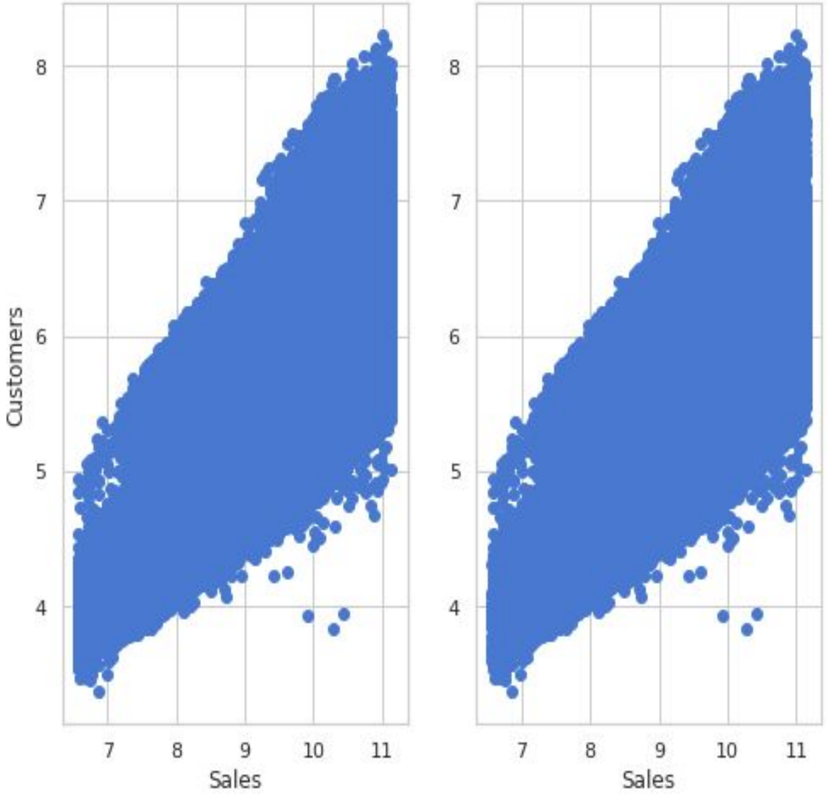
- Date:- Extracted Day, Month,Year & Drop the Date Column Itself.
- StateHoliday:-In the StateHoliday We have a,b,c as Holidays hence replced these with 1 which represent Holidays and 0 as working day.Then Made the column into integer type.
- StoreType:-It has Nominal kind of data so Encoded with Dummies.
- Assortment:-It also has Nominal Kind of data so encoded with Dummies using one Hot Encoding.

Handling Missing Values:

- There were missing values in six Features of Store Data set.
- CompetitionOpenSinceMonth & CompetitionOpenSinceYear : Here Imputed the missing values with respective mode values.
- competitionDistance: Imputed the missing values with median as there are few Outliers in the feature.
- Promo2SinceWeek ,Promo2SinceYear : Imputed missing with respective median values.(Mode can distort the Feature data as the missing values were very high in number.)
- PromoInterval:Encoded with label then converted to string and then imputed with median.(Since it had lots of Nulls).

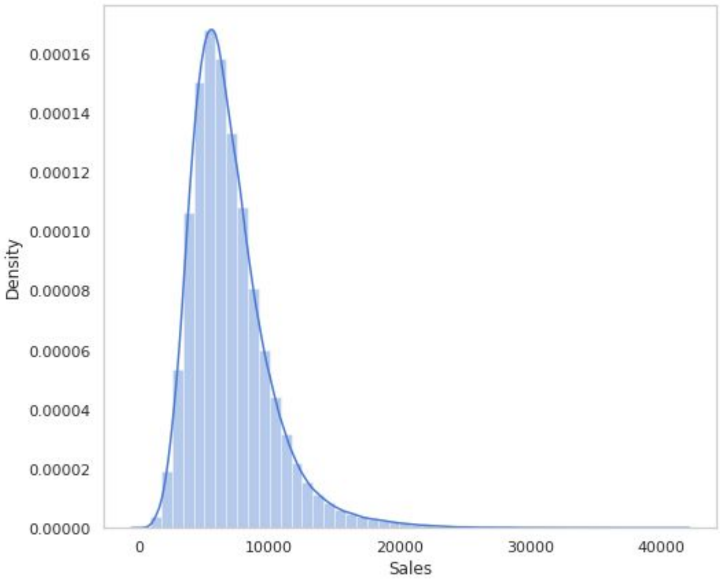
Handling Outliers

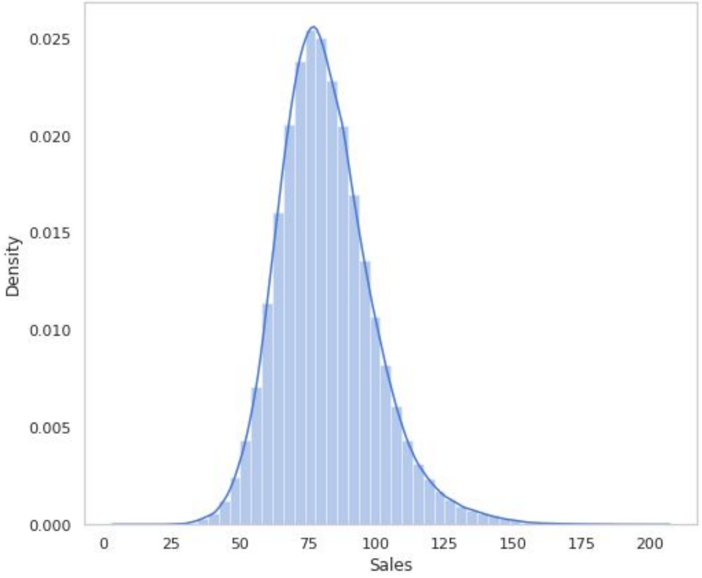
- Sales(Target Variable):-Removed all the outliers from the target variable.
- Customers:-While removing outliers from target variable we also end up removing outliers from customers columns.



Exploratory Data Analysis

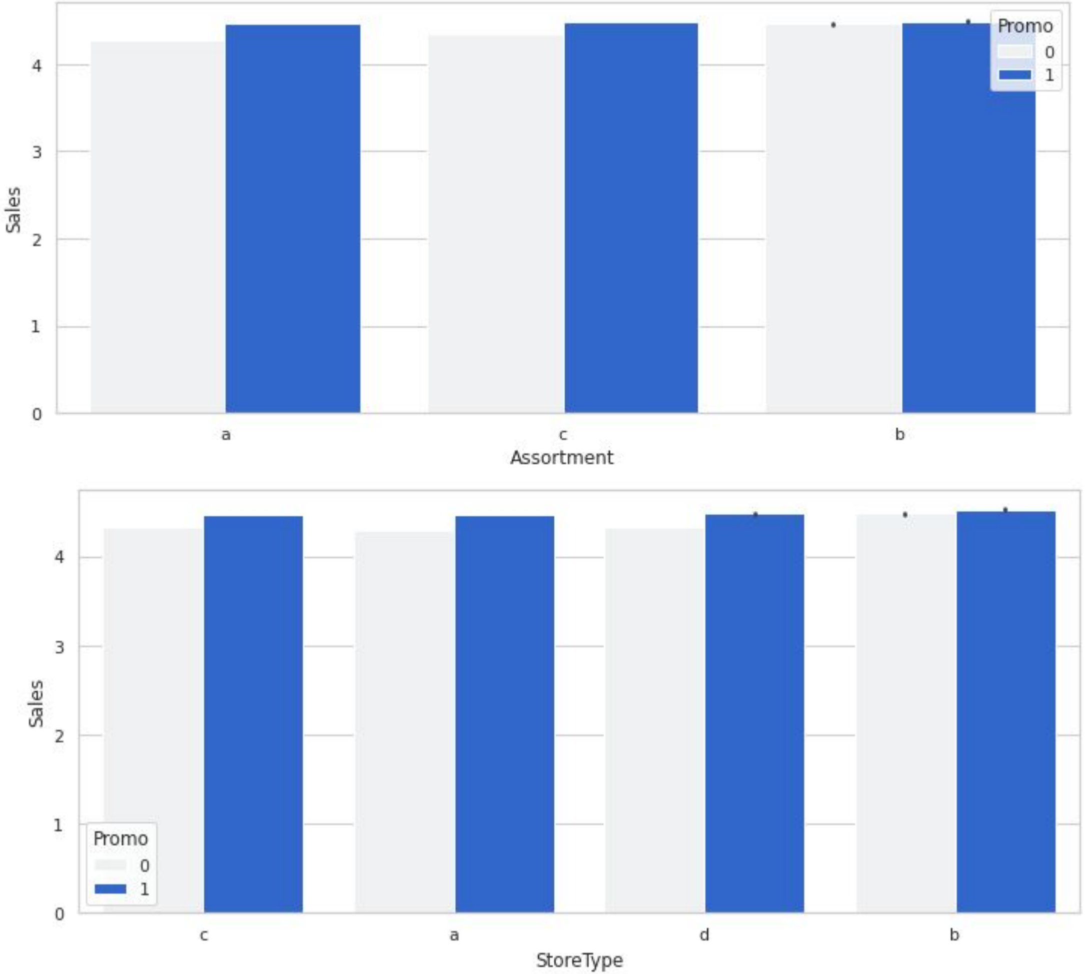
Improving Target Variable



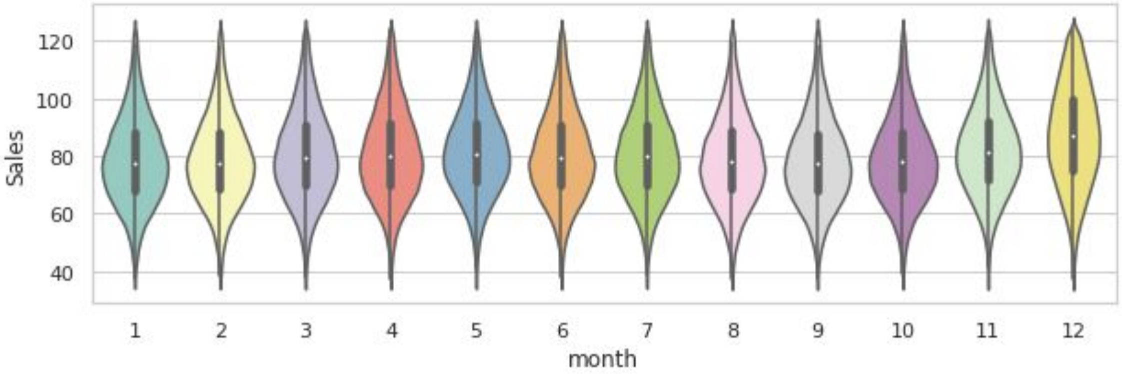


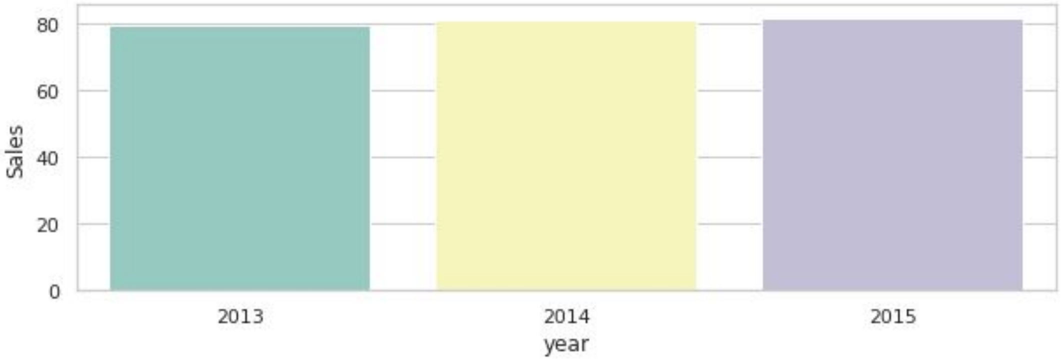
For Improving target Variable we have removed large amount and zeros and transformed sales into their normal distribution from slight left distribution.

StoreType and Assortment with promo vs Sales.

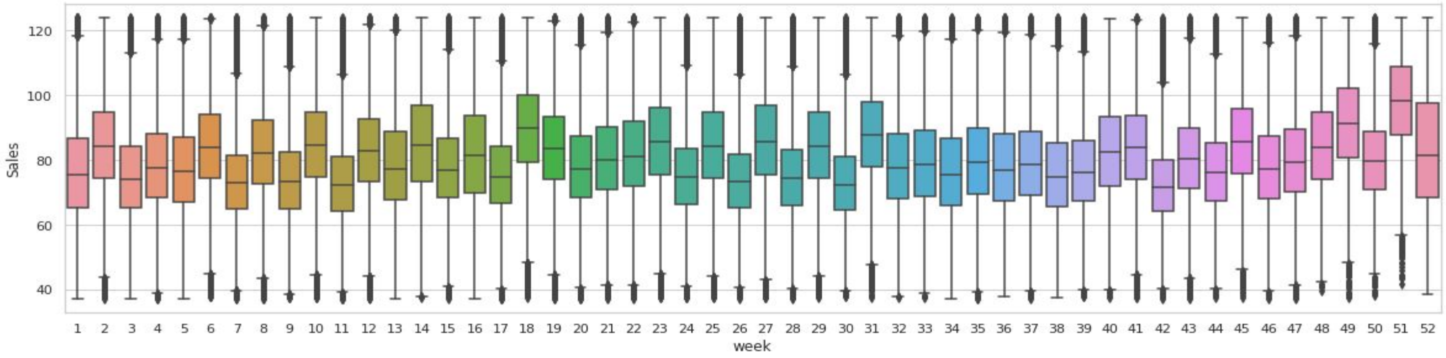


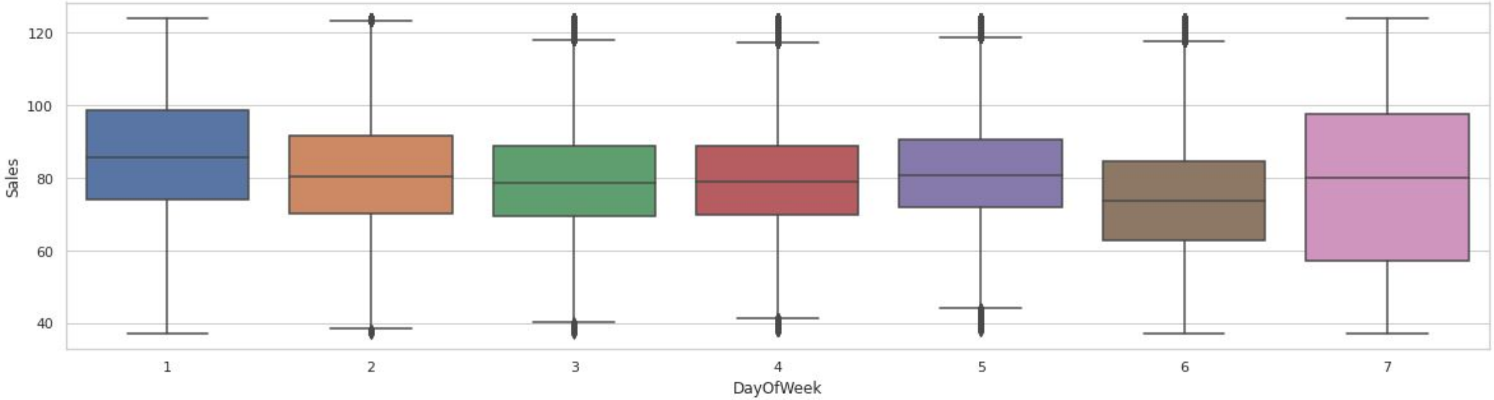
Sales Basis of year and month





Sales Basis on week and day

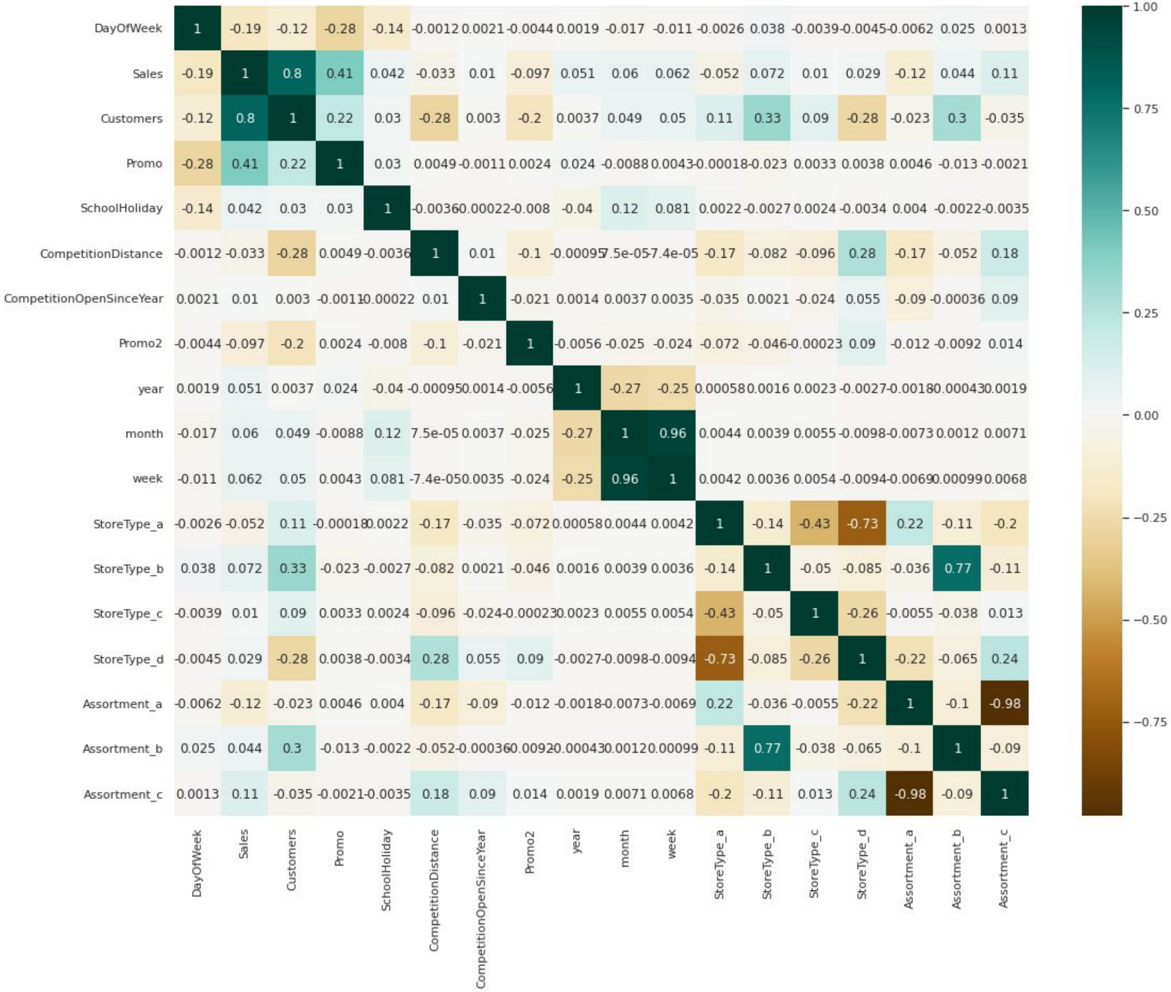




Correlation:

Conclusion:

- For correlation of rows where the dependent variable (Sales in this case) is not involved because if a variable is correlated with the dependent variable then this would be a good sign for our model.
- Correlation within dependent variables is what we need to look for and avoid. This data doesn't contain perfect multicollinearity among independent variables.



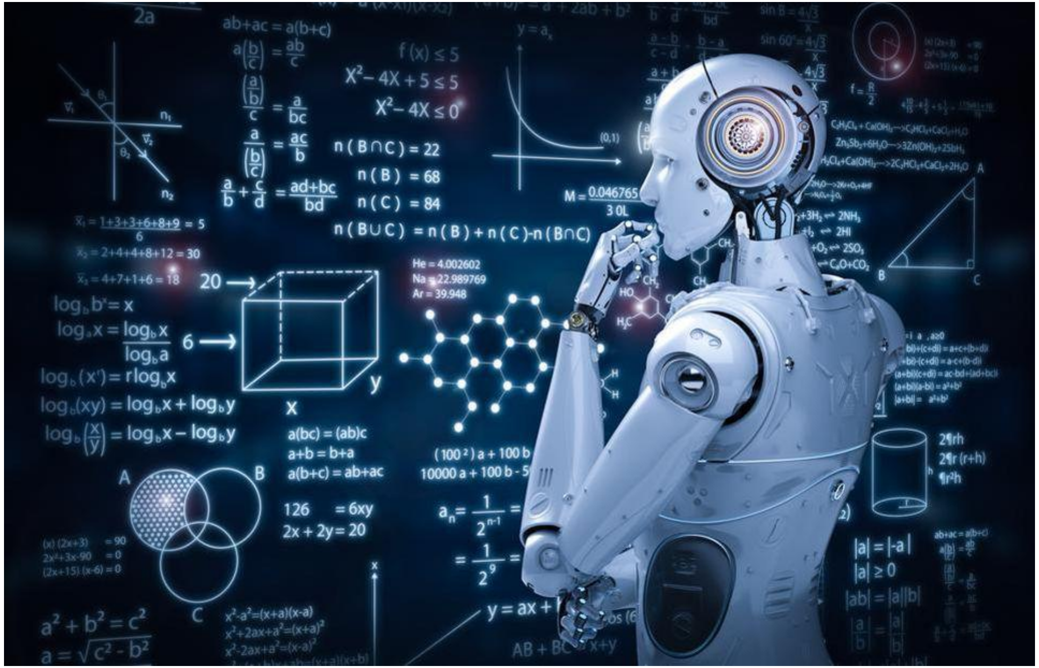
	DayOfWeek	Sales	Customers	Promo	SchoolHoliday	CompetitionDistance	CompetitionOpenSinceYear	Promo2	year	month	week	StoreType_a	StoreType_b	StoreType_c	StoreType_d	Assortment_a	Assortment_b	Assortment_c
DayOfWeek	1	-0.19	-0.12	-0.28	-0.14	-0.0012	0.0021	-0.0044	0.0019	-0.017	-0.011	-0.0026	0.038	-0.0039	-0.0045	-0.0062	0.025	0.0013
Sales	-0.19	1	0.8	0.41	0.042	-0.033	0.01	-0.097	0.051	0.06	0.062	-0.052	0.072	0.01	0.029	-0.12	0.044	0.11
Customers	-0.12	0.8	1	0.22	0.03	-0.28	0.003	-0.2	0.0037	0.049	0.05	0.11	0.33	0.09	-0.28	-0.023	0.3	-0.035
Promo	-0.28	0.41	0.22	1	0.03	0.0049	-0.0011	0.0024	0.024	-0.0088	0.0043	-0.00018	-0.023	0.0033	0.0038	0.0046	-0.013	-0.0021
SchoolHoliday	-0.14	0.042	0.03	0.03	1	-0.0036	0.00022	-0.008	-0.04	0.12	0.081	0.0022	-0.0027	0.0024	-0.0034	0.004	-0.0022	-0.0035
CompetitionDistance	-0.0012	-0.033	-0.28	0.0049	-0.0036	1	0.01	-0.1	-0.00095	7.5e-05	-7.4e-05	-0.17	-0.082	-0.096	0.28	-0.17	-0.052	0.18
CompetitionOpenSinceYear	0.0021	0.01	0.003	-0.0011	0.00022	0.01	1	-0.021	0.0014	0.0037	0.0035	-0.035	0.0021	-0.024	0.055	-0.09	-0.00036	0.09
Promo2	-0.0044	-0.097	-0.2	0.0024	-0.008	-0.1	-0.021	1	-0.0056	-0.025	-0.024	-0.072	-0.046	-0.00023	0.09	-0.012	-0.0092	0.014
year	0.0019	0.051	0.0037	0.024	-0.04	-0.00095	0.0014	-0.0056	1	-0.27	-0.25	0.00058	0.0016	0.0023	-0.0027	-0.0018	0.00043	0.0019
month	-0.017	0.06	0.049	-0.0088	0.12	7.5e-05	0.0037	-0.025	-0.27	1	0.96	0.0044	0.0039	0.0055	-0.0098	-0.0073	0.0012	0.0071
week	-0.011	0.062	0.05	0.0043	0.081	-7.4e-05	0.0035	-0.024	-0.25	0.96	1	0.0042	0.0036	0.0054	-0.0094	-0.0069	0.00099	0.0068
StoreType_a	-0.0026	-0.052	0.11	-0.00018	0.0022	-0.17	-0.035	-0.072	0.00058	0.0044	0.0042	1	-0.14	-0.43	-0.73	0.22	-0.11	-0.2
StoreType_b	0.038	0.072	0.33	-0.023	-0.0027	-0.082	0.0021	-0.046	0.0016	0.0039	0.0036	-0.14	1	-0.05	-0.085	-0.036	0.77	-0.11
StoreType_c	-0.0039	0.01	0.09	0.0033	0.0024	-0.096	-0.024	-0.00023	0.0023	0.0055	0.0054	-0.43	-0.05	1	-0.26	-0.0055	-0.038	0.013
StoreType_d	-0.0045	0.029	-0.28	0.0038	-0.0034	0.28	0.055	0.09	-0.0027	-0.0098	-0.0094	-0.73	-0.085	-0.26	1	-0.22	-0.065	0.24
Assortment_a	-0.0062	-0.12	-0.023	0.0046	0.004	-0.17	-0.09	-0.012	-0.0018	-0.0073	-0.0069	0.22	-0.036	-0.0055	-0.22	1	-0.1	-0.98
Assortment_b	0.025	0.044	0.3	-0.013	-0.0022	-0.052	-0.00036	-0.0092	0.00043	0.0012	0.00099	-0.11	0.77	-0.038	-0.065	-0.1	1	-0.09
Assortment_c	0.0013	0.11	-0.035	-0.0021	-0.0035	0.18	0.09	0.014	0.0019	0.0071	0.0068	-0.2	-0.11	0.013	0.24	-0.98	-0.09	1

Feature Selection

- After performing Feature Engineering, EDA and Fitting Our Features to Assumptions we come to conclusion over which Feature to select and which to discard for our upcoming Models.
- We Discarded Promo as they have huge nulls imputing those can manipulating our final conclusion.

Models Performed:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic net
- Decision Tree Regressor
- Random Forest



Model Observations with Evaluation Matrice

Linear Regression Model with Cross_Validation:-

1.Our Data set is not perfect suitable for the linear regression model.

2.We tried cross validation with linear regression but still model improve by 0.01 percent.

Lasso and Ridge Regression with Cross_validation & HyperParameter:-

They both also unable to improve the accuracy of our model even after HyperParameter tunning.

Decision Tree Regression:-

It works Excellent on our Data set. Giving R2 of 0.99. We can also say the model has overfitted our dataset for the Decision Tree Regression.

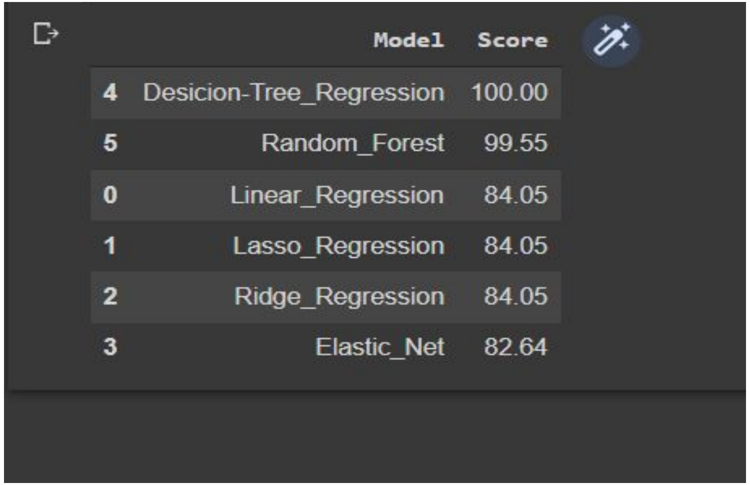
Random Forest Regression:-

It also works great on our Data set. Giving R2 of 0.99.

Model Comparison

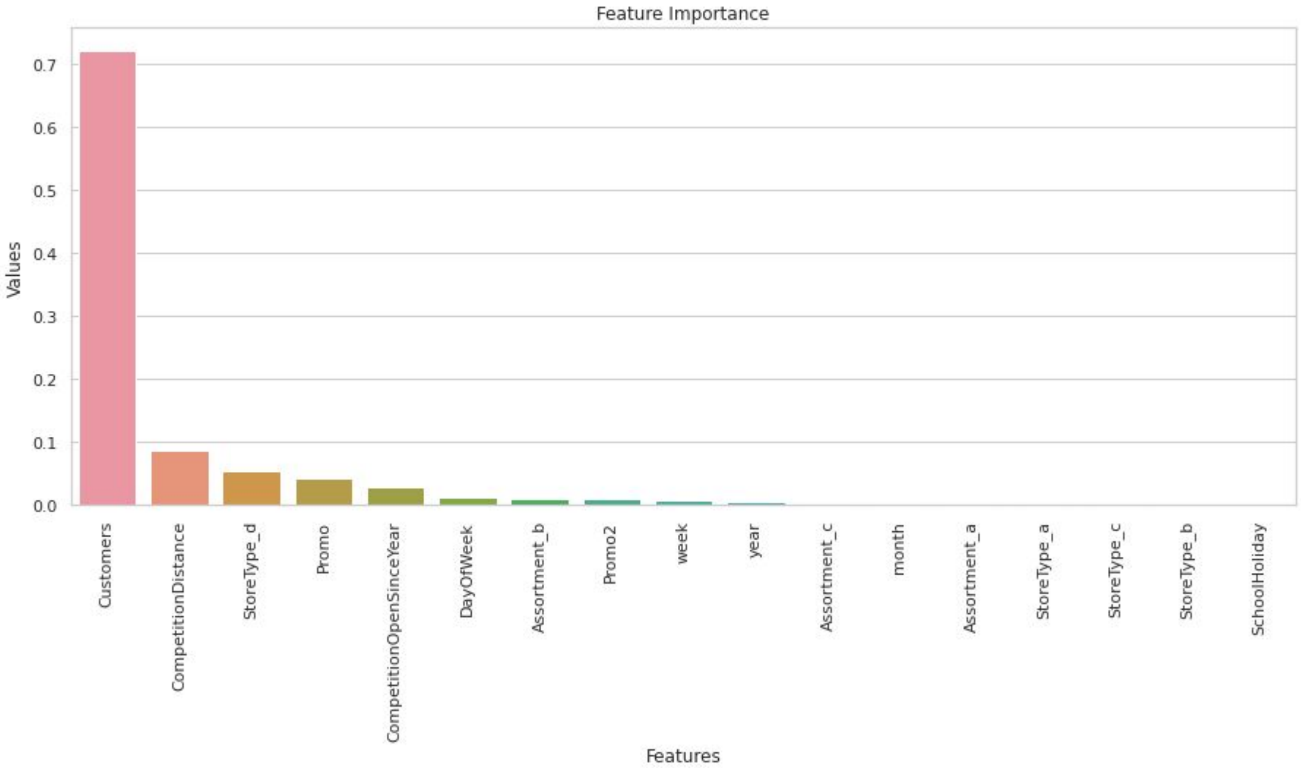
84.05

3 Elastic_Net



	Model	Score
4	Desicion-Tree_Regression	100.00
5	Random_Forest	99.55
0	Linear_Regression	84.05
1	Lasso_Regression	84.05
2	Ridge_Regression	84.05
3	Elastic_Net	82.64

Important Features



Challenges Faced:

1.The Data set was comparatively very large hence computation time was also more than expected.

2. There were lots of Outliers in our Target variable Handling them was difficult decision as removing them could harm our data and keeping them could distort our model.We choose to remove them as our Data set big.

3.The Features were mostly categorical type hence we have faced lot of challenges to fit it into regression type model.

4.Decision tree regression and random forest both takes lot's of time to execute as they heavy computation in nature.

5.Due to lot's of complex operations there can be loss of data sometimes ,So constant saving of datasets and operations is required.

6. It takes constant editing of datasets which can be time consuming and complex sometimes.

Conclusion

- Sundays have negligible sales records. Monday(1) and Friday(5) have highest sales. Fridays have maximum sales records. Customer feature also follow same trend.
- Lots of zero sales is disturbing our Target Variable. When stores are closed the sales value is zero hence we have deal with the zero sales.
- During State Holidays there is negligible sales records.But we have some Sales records even during School Holidays.
- Sales has declined in year 2015 compare to previous years.
- We Tried to fit our dataset into regression assumption but we fail to fit most of the features due to their nature.

- From the model Comparison we can say our dataset is not performing well in linear Regression even after Cross validation , Hyperparameter tuning with L1 & L2 does not help the model to improve accuracy its accuracy.
- Decision Tree and Random Forest both working well. It should be due the nature of features which mostly categorical in nature.
- But it also seem that they are overfitting .
- From the Model we come to know that customers feature is one of the most important features as it should be in case of sales prediction.
- Competition Distance , Competition Since year, Days of week all these features are also important for the sales prediction.

Thank you