RETAIL SALES PREDICTION

Raja Chowdhury
Data Science Trainees
Alma Better, Bangalore

Abstract:

In this project, we applied machine learning techniques to a real-world problem of predicting stores sales. This kind of prediction enables store managers to create effective staff schedules that increase productivity and motivation. We used popular open-source statistical programming language Python. This case study solves everything right from scratch. So, you will get to see each and every phase of how in the real world a case study is solved.

Problem Statement

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.

You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

Introduction

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. 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.

In 2019 Rossmann had more than €10 billion turnover in Germany, Poland, Hungary, the Czech Republic, Turkey, Albania, Kosovo and Spain. In 2021, sales increased by 8.1 percent to 11.1 billion euros. There is a total of 4,361 Rossmann branches, 2,231 of which are in Germany.

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. There is also a perfume range with around 200 commercial brands. Rossmann has 29 private brands with 4600 products (as of 2019). In 1997, the first own brands Babydream, Facelle, Sunozon and Winston were introduced. The best-

known Rossmann brands are Isana (skin, hair and body care), Alterra (natural cosmetics), domol (cleaning and laundry detergents) alouette (paper tissues etc).

Our Approach:

1. Steps performed in this ML Supervised Project

Handling dataset with the fundamental steps to unveil the factors:

- Importing Libraries and Loading the Datasets
- Overview of The Datasets
    - Reading & Inspection of First Dataset
    - Reading & Inspection of Second Dataset
    - Merging both the datasets
    - Further analysing both the datasets
- Data Wrangling and Processing
    - Changing Data Types of Columns
    - Extracting Date
    - Combining and Creating Columns
    - Null Value Treatment
    - Handling Outliers
- Exploratory Data Analysis
- Key Findings From EDA
- Feature Engineering
    - Feature Selection
    - Correlation
    - Square Root Transformation
    - Scaling Numerical Columns
    - Dummification
- ML Model
    - Train-Test Split
    - Model Training and Prediction
        - Linear Regression
        - Lasso
        - Ridge
        - Decision Tree
        - Random Forest
        - Gradient Boosting
        - XGBoost
    - Hyperparameter Tuning on All Regression Model
    - Cross Validation
    - Feature Importance
- Conclusion

2. Data Understanding

As the objective is clear the data needs to be analysed and this process starts with understa nding our dataset of Rossmann Stores. Our dataset contains the following variables:

- Id - an Id that represents a (Store, Date) duple within the test set
- Store (nominal)- a unique Id for each store
- Sales (discrete)- the turnover for any given day (this is what you are predicting)
- Customers(discrete) - the number of customers on a given day
- Open(nominal) - an indicator for whether the store was open: 0 = closed, 1 = open
- State Holiday(nominal) - 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(nominal) - indicates if the (Store, Date) was affected by the closure of public schools
- Store Type(nominal) - differentiates between 4 different store models: a, b, c, d
- Assortment(nominal) - describes an assortment level: a = basic, b = extra, c = extended
- Competition Distance(continuous) - distance in meters to the nearest competitor store
- CompetitionOpenSince [Month/Year] (discrete) - gives the approximate year and month of the time the nearest competitor was opened
- Promo(nominal) - indicates whether a store is running a promo on that day
- Promo2(nominal) - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- Promo2Since [Year/Week] (discrete) - describes the year and calendar week when the store started participating in Promo2
- Promo Interval(discrete) - 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
- DayOfWeek (ordinal) - Day of the week, using 1-7 for Monday - Sunday respectively
- Date (Date) - Date of the entry
- PromoOpen- Time from when store is participating in the promo
- CompetitionOpen- Time from when competitor arrived

3. Data Processing & Preparation

- CompetitionOpenSinceMonth, CompetitionOpenSinceYear, Promo2SinceWeek, and Promo2SinceYear columns are only using whole numbers and they are a discrete value, so we will change them from floats to integers.
- Changed the data type of Date column from object type to Datetime.
- we have created new column called "PromoOpen" from existing columns to measure more accurate period (in Months) from when the store was participating in the Promo2.
- PromoOpen column has figures in negative which indicates that the store has not started any promo yet. So, we have to replace those negative figures with zero.

- we have created new column called "Competition Open" from existing columns to measure the period (in months) from when the nearest competition has opened.

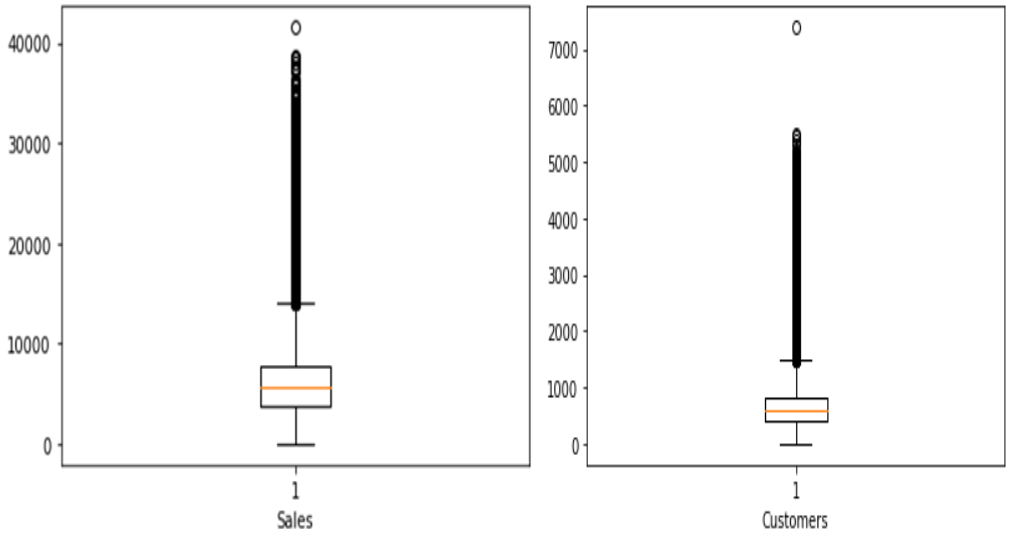
- Competition Open column has figures in negative which indicates that the store does not have any competition. So, we have to replace those negative figures with zero.

Null Value Treatment

- Promo Interval and PromoOpen has similar number of null values which means when promo is not open then promo have no interval as well.
- We can see that whenever the store was not participating in Promo2, we had null values present in PromoOpen and Promo Interval columns as well. we have imputed zero in place of null values present in PromoOpen and Promo Interval columns because logically when promo2 is zero then PromoOpen and Promo Interval should be zero too.

Handling Outliers

- We used boxplot to detect outliers.
- We removed outliers using z score method on sales and customer column.
- After that we have replaced other outliers with different percentile values using capping method as the number of outliers were huge.



4. Data Analysis and Findings

- **Which store model has the maximum sales?**

  From the first graph it is clear that 'a' Store Model have the maximum number of sales and store counts followed by 'd' while Store Model 'b' have the least number of sales and store counts.

- **What is the average sales and customer from the type of store models?**

  From second graph it is surprising to see that store model "b" which have least number of store counts performs quite well on average sales and customers compared to other store models.

- **Which assortment level has maximum number of sales and store counts**

  We can see from the graph that Basic Assortment level have the maximum number of sales and store counts followed by Extended level while Extra Assortment have the least number of sales and store counts

  We can infer from the graph that assortment level 'b' with least store counts have perform quite well compared to 'a'. While there is another surprising fact that assortment level 'c' has maximum number of sales with the least number of customers.

- **Is there any relationship between customers and sales?**

  Here we can see from the graph that there is a linear relationship between customer and sales and it is also noticeable that whenever promo was open, the store has higher sales and customer compared to the period when promo was closed, which means promo had good impact on the sales.

- **At what time of the year has more sales is highest and lowest?**

We can infer from the graph that the sales are highest on 30th followed by 2nd and 4th date of every month while sales are lowest on the 1st date of every month followed by 25th and 26th date.

Here we can infer from the graph that the sales are at maximum on Mondays while sales are zero on Sunday because it seems like store use to remain closed on Sundays.

Here we can see that the Sales and Customers are at peak during November and December due to festive season like Christmas while sales are at lowest during January and May or we may say these months to be off season.

Here we can see that during public holidays store made more sales compared to Easter and Christmas holidays.

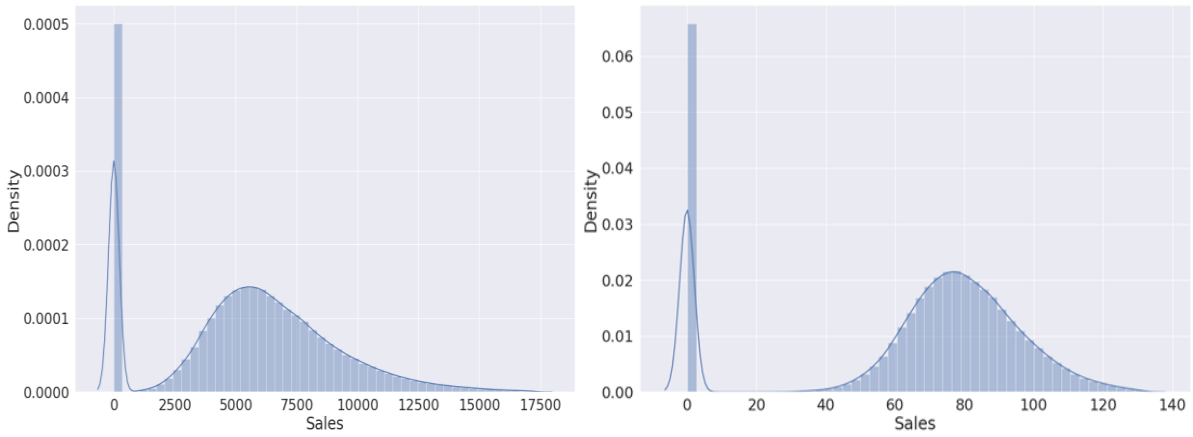
- **Does school affect any kind of sales?**

  Here we can see that 17.8% of the sales gets affected by the school holidays which also means that around 17% of the sales are oriented from the school students.

5. Feature Engineering

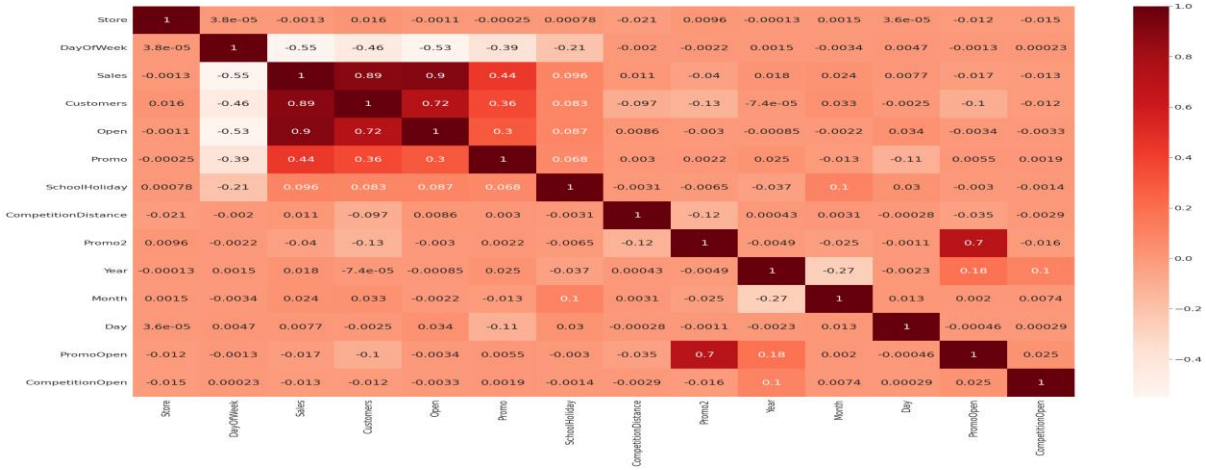
Feature Selection: We dropped Store column to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model, as it was irrelevant variable.

Dependent Variable transformation: We have noticed that our sales column was right skewed so here we have used square root transformation to remove the skewness. Transforming our dependent variable results in a different interpretation of our coefficient values.

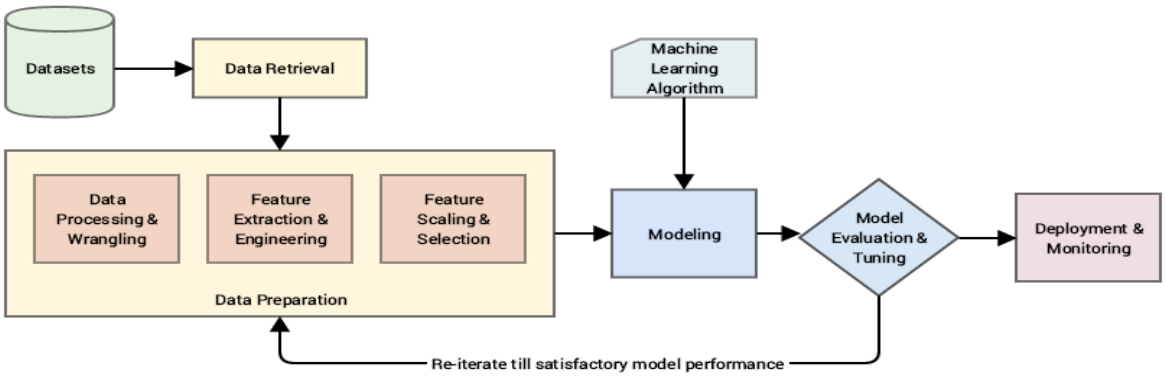


Scaling Numerical Columns: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range so We have scaled numerical columns and now all values lie between 0 and 1.

Correlation: Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A but we didn't find any strong correlation between independent variables but we found some correlation with our dependent feature which is a good sign for our model



Dummification: Dummy or Boolean variables are qualitative variables that can only take the value 0 or 1 to indicate the absence or presence of a specified condition so we used this on categorical variables.



7. Machine Learning Model

After doing train test split, we used 7 algorithms to train and test our model.

- **Linear Regression -** Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
- **Lasso Regression -** Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The

lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

- **Ridge Regression-** Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values
- **Decision Tree-** Decision tree builds 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.
- **Random Forest**- Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.
- **Gradient Boosting-** Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.
- **XG boost-** XGBoost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XGBoost) objective function and base learners. The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e., how far the model results are from the real values. The most common loss functions in XGBoost for regression problems is reg: linear, and that for binary classification is reg: logistics.

Evaluation metrics used-

- **Mean Squared Error-** The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.
- **Root mean squared error-** Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far

from the regression line data points are; RMSE is a measure of how to spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

- **R2-** R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).
- **Mean Absolute Error-** is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set.
- **Adjusted R2-** The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not.
- **Root mean squared percentage error-** One can compute the ratio of the MAE or RMSE to the mean of the target variable to get a percentage error for the validation data. Alternatively, one can directly compute the MAPE or RMSPE over the validation data and get the percentage error that way.

Hyperparameter Tuning

Hyperparameters are the knobs or settings that can be tuned before running a training job to control the behaviour of an ML algorithm. They can have a big impact on model training as it relates to training time, infrastructure resource requirements (and as a result cost), model convergence and model accuracy. Model parameters are learnt as part of training process, whereas the values of hyperparameters are set before running the training job and they do not change during the training.

After using hyperparameter tuning on all our regression models we came to our final conclusion that **Random Forest Regression** have performed much better with Adjusted R2 : 0.99409 as Random Forest Regression can handle large datasets efficiently and the random forest algorithm provides a higher level of accuracy in predicting outcomes over any other regression algorithm.

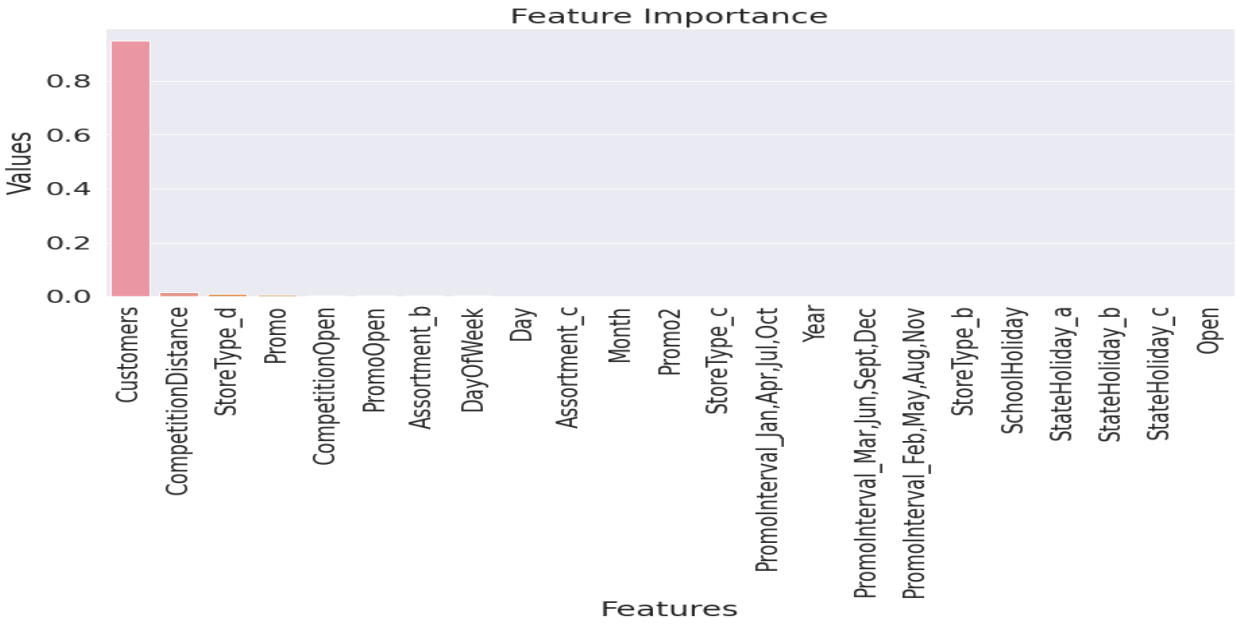
Cross Validation

Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, ie, failing to generalize a pattern. the k-fold cross-validation method to perform cross-validation. In k-fold cross-validation, you split the input data into k subsets of data (also known as folds). You train an ML model on all but one (k-1) of the subsets, and then evaluate the model on the subset that was not used for training. This process is repeated k times, with a different subset reserved for evaluation (and excluded from training) each time.

Hyper-Tuned Regression Techniques	MAE	MSE	RMSE	R2	Adjusted R2
Lasso	5.075486	42.034368	6.483392	0.96282	0.962816
Ridge	5.075483	42.034367	6.483392	0.96282	0.962816
Decision Tree	2.154891	11.09179	3.330434	0.990189	0.990188
Random Forest	**1.706587**	**6.680834**	**2.584731**	**0.994091**	**0.99409**
Gradient Boosting	3.308434	19.519283	4.418063	0.982735	0.982733
XGboost	3.493811	22.214871	4.713265	0.980351	0.980349

Feature Importance:

This technique calculates a score for all the input features for a given model the scores simply represent the "importance" of each feature. A higher score means that the specific feature will have a larger effect on the model that is being used to predict a certain variable. Our model and prediction are highly influence by the number of customers.



8. Challenges Faced:

- Handling and understanding large amount of data (1017209 number of records and 18 number of fields)
- Columns with improper data type and wrong values.
- Combining, creating and removing columns.
- Records containing more than 50% of nan values and replacing it with substitutes.
- Removing and replacing outliers from dependent and independent variables.
- Reducing skewness from the variables.
- Feature selections for ML Model.
- Converting columns with categorical variables to integer type and scaling numerical variables for regression models.
- Performing and choosing right kind of model.

9. Conclusions

1. Store model 'b' have least number of stores in Rossmann yet it performed well and made more sales than other store models so it is advisable to increase the number of 'b' store model.
2. Assortment level 'Basic' have the maximum number of stores in Rossmann yet it performed very badly but at the same time 'Extra' and 'Extended' assortment level with a smaller number of stores had performed extra ordinarily so it would be advisable to increase these assortment level.
3. Linear relationship has been found among customer, sales and promo. And it has been seen that most of the customers came for shopping during the promo days as the cost was lower on those days. So, promo should be initiated to more stores to increase the sales.
4. Sales has been low on the initial days of the month as compared to the end days; it can be assumed that people used to shop for the next month at the end of the previous month. Those products can be mainly be of basic necessities of a person's daily life.
5. Average sales on weekdays were more as compared to weekends because promos were provided to the customers during weekdays to increase the sales and not to weekends and reason might be that store use to remain close on Sundays.
6. Sales during November and December month was higher compared to other months and that can be due to festive season in western European countries.
7. Mostly competitor stores weren't that far and the stores were densely located near each other and also sales were higher when competition was nearer. So, when competition was not there and sales were also low that might be because of the location factor like rural area.

8. School holidays also influence the sales a lot as it can be observed that 17.8% of the sales gets affected by the school holidays which also means that around 17% of the sales are oriented from the school students.
9. Performing various regression techniques, we found that XGboost Regression model have the better performance (with R2 : 0.988409) but after applying hyperparameter tuning on all our models we finally came to the conclusion that Random Forest Regression model have even higher performance (with R2 :0.994091) among the other models, as Random Forest Regression can handle large datasets efficiently and it's algorithm provides a higher level of accuracy in predicting outcomes over any other regression algorithm.
10. According to cross validate method of sklearn while cross validating 3 times our random forest regression model had the following performance:

    Average fit_time is 458.28843

    Average score_time is 18.84081

    Average R2 score is 0.99338