**Retail Sales Prediction**

Nitesh Bhowmick,

Ayush Goyal, M Sameer Ahmed,

Data science trainees

Alma Better, Bangalore

**Abstract:**

The study conducted was on the proposed “Retail Sales Prediction”. The main purpose of this study was to create a program that will predict the Retail Sales according to features. The researchers used the descriptive type of survey methods where they distributed questionnaires to the respondent of the study as a research instrument for data gathering.

Our experiment can help understand what could be the reason for the regression of such labels by feature selection, data analysis and prediction with Python Programming & ML algorithm taking into account previous trends to determine the correct classification.

**1. Problem Statement**

This data set contains information for historical data including Sales & supplemental information about the stores. The data fields are Id, Store, Sales, Customers, Open, State Holiday, School Holiday, Store type, Assortment, Competition Distance, CompetitionOpenSince[Month/Year], Promo, Promo 2, Promo2Since[Year/Week], PromoInterval. All personally identifying information has from the data. We will perform exploratory data analysis with python to get insight from the data & using ML

algorithm to get the accuracy level of all algorithms.

**2. Introduction**

### 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.

I will proceed with reading the data, and then perform data analysis. The practice of examining data using analytical or statistical methods in order to identify meaningful information is known as data analysis. After data analysis, we will find out the data distribution and data types. We will train 3 regression ML algorithms to predict the output. We will also compare the outputs.

Let us get started with the project implementation.

### **The Rossmann Sales Company**

### A private drug store chain based in Germany, with main operations in Europe. Operates over 3,000 drug stores in 7 different countries.

### Offers healthcare and beauty products, including baby and body care, hygiene, cosmetics, dental hygiene, hair care, and so on.

### Business Model: Product sales.

### **Problem**

### The CFO wanted to reinvest in all stores, therefore, he needs to know how much revenue each store will bring so he can invest it now.

### **Goal**

### Predict the daily sales of all stores for up to six weeks in advance.

### **Deliverables**

### Model's performance and results report with the following topics:

### What are the daily sales in dollars for the next 6 weeks?

### Predictions will be available through a Telegram Bot where stakeholders can access the prediction by a smartphone

**3. Key features of Retail Sales Prediction:-**

### Most of the fields are self-explanatory. The following are descriptions for those that aren't.

* **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 has 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, and November of any given year for that store

## **4. Methodology**

This dataset is a live dataset of Roseman Stores. In analysing this problem we observe that the Roseman problem is a regression problem and our primary goal is to predict the sales figures of the Roseman problem. In this Notebook, we work on the following topics

Analysing the Dataset by using Exploratory Data Analysis. Using Exponential Moving Averages analyse Trends and Seasonality in the Roseman dataset. Analyse Regression analysis using the following prediction analysis, A. Linear Regression Analysis B. Elastic Regression ( Lasso and Ridge Regression). C. Decision Tree

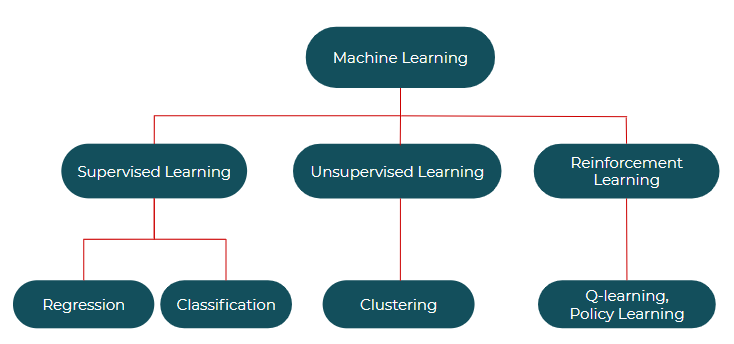
### **5. Machine Learning:**

**Machine Learning:**

Learning is any process by which a system improves performance from experience.

Machine Learning is concerned with computer programs that automatically improve their performance through experience.

**Task in Machine Learning:-**



**Supervised learning: -** Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data.

**Types:-**

* Linear Regression
* Logistic Regression
* Classification
* Naive Bayes Classifiers
* K-NN (k nearest neighbours)
* Decision Trees
* Support Vector Machine

**Advantages:-**

* Supervised learning allows collecting data and produces data output from previous experiences.
* Helps to optimize performance criteria with the help of experience.
* Supervised machine learning helps to solve various types of real-world computation problems.

**Disadvantages:-**

* Classifying big data can be challenging.
* Training for supervised learning needs a lot of computation time. So, it requires a lot of time.

#### **Unsupervised learning:-**

Unsupervised learning is the training of a machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided which means no training will be given to the machine. Therefore the machine is restricted to finding the hidden structure in unlabeled data by itself.

**Types of Unsupervised Learning:-**

**Clustering:-**

1. Exclusive (partitioning)
2. Agglomerative
3. Overlapping
4. Probabilistic

**Clustering Types:-**

1. Hierarchical clustering
2. K-means clustering
3. Principal Component Analysis
4. Singular Value Decomposition
5. Independent Component Analysis

**6. Steps involved:**

* **Exploratory Data Analysis**

 In statistics, exploratory data analysis (*EDA*) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily *EDA* is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.

* **Null values Treatment**

Our dataset contains nan values in-store data. There were a lot of nan values in some columns. that had to be dealt with. So I wrote a code to specifically deal with the nan values of each column either by replacing it with 0, mode or median. Removing the columns was not an option as they might remove some significant amount of data.

* **Fitting different models**

For Analysis we tried various types of ML models:

1. **Linear Regression.**
2. **Decision Tree.**
3. **Lasso.**

# **Ridge.**

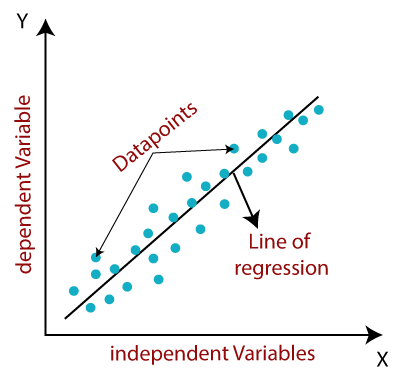
**7. Algorithms:**

1. **Linear Regression:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as :

**y= a0+a1x+ ε**

Here,

**Y**= Dependent Variable (Target Variable)

**X**= Independent Variable (predictor Variable)

**a0**= intercept of the line (Gives an additional degree of freedom)

**a1** = Linear regression coefficient (scale factor to each input value).

**ε** = random error

The values for x and y variables are training datasets for Linear Regression model representation.

## **Types of Linear Regression:-**

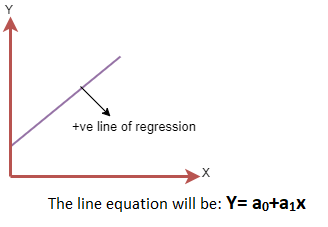
Linear regression can be further divided into two types of algorithms:

* **Simple Linear Regression:**  
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
* **Multiple Linear regression:**  
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

## **Linear Regression Line:-**

A linear line showing the relationship between the dependent and independent variables is called a regression line. A regression line can show two types of relationship:

* **Positive Linear Relationship:**  
  If the dependent variable increases on the Y-axis and the independent variable increases on X-axis, then such a relationship is termed a Positive linear relationship.



* **Negative Linear Relationship:**  
  If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

**Finding the best fit line:**

When working with linear regression, our main goal is to find the best fit line which means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a0, a1) give a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use the cost function.

### **Cost function:**

* The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
* Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
* We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.

For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average squared error that occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

Linear Regression in Machine Learning

Where

**N**=Total number of observations

**Yi** = Actual value

**(a1xi+a0)**= Predicted value.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so the cost function will be high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

* **Import Linear Regression:-**

from sklearn.linear\_model import LinearRegression

1. **Decision Tree:-**

# A decision tree is a support tool with a tree-like structure that models probable outcomes, cost of resources, utilities, and possible consequences. Decision trees provide a way to present [algorithms](https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/what-are-algorithms-algos/) with conditional control statements. They include branches that represent decision-making steps that can lead to a favorable result.

# Decision trees are one of the best forms of learning algorithms based on various learning methods. They boost predictive models with accuracy, ease in interpretation, and stability. The tools are also effective in fitting non-linear relationships since they can solve data-fitting challenges, such as regression and classifications.

# **Terminology:-**

* **Root Node**: It represents the entire population or sample and this further gets divided into two sets.
* **Splitting**: It is a process of dividing a node into two sub-nodes.
* **Decision Node:** When a sub-node splits into further sub-nodes, then it is called a decision node.
* **Leaf/ Terminal Node:** Nodes that do not split is called leaf or terminal node.
* **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
* **Branch / Sub-Tree:** A sub-section of an entire tree is called a branch or sub-tree.
* **Parent and Child Node:** A node, which is divided into sub-nodes is called the parent node of sub-nodes where as sub-nodes are the child of the parent node.

### **Types of Decisions**

There are two main types of decision trees that are based on the target variable, i.e., categorical variable decision trees and continuous variable decision trees.

#### **1. Categorical variable decision tree**

A categorical variable decision tree includes categorical target variables that are divided into categories. For example, the categories can be yes or no. The categories mean that every stage of the decision process falls into one category, and there are no in-betweens.

#### **2. Continuous variable decision tree**

A continuous variable decision tree is a decision tree with a continuous target variable. For example, the income of an individual whose income is unknown can be predicted based on available information such as their occupation, age, and other continuous variables.

### **Applications of Decision Trees:-**

#### **1.** Assessing prospective growth opportunities

One of the applications of decision trees involves evaluating prospective growth opportunities for businesses based on historical data. Historical data on sales can be used in decision trees that may lead to making radical changes in the strategy of a business to help aid expansion and growth.

#### **2.** Using demographic data to find prospective clients

Another application of decision trees is in the use of [demographic data](https://corporatefinanceinstitute.com/resources/knowledge/economics/demographics/) to find prospective clients. They can help streamline a marketing budget and make informed decisions on the target market that the business is focused on. In the absence of decision trees, the business may spend its marketing market without a specific demographic in mind, which will affect its overall revenues.

#### **3.** Serving as a support tool in several fields

Lenders also use decision trees to predict the probability of a customer defaulting on a loan by applying predictive model generation using the client’s past data. The use of a decision tree support tool can help lenders evaluate a customer’s creditworthiness to prevent losses.

Decision trees can also be used in operations research in planning logistics and [strategic management](https://corporatefinanceinstitute.com/resources/knowledge/strategy/strategic-management/). They can help in determining appropriate strategies that will help a company achieve its intended goals. Other fields where decision trees can be applied include engineering, education, law, business, healthcare, and finance.

**Advantages of Decision Trees**

#### **1. Easy to read and interpret**

One of the advantages of decision trees is that their outputs are easy to read and interpret without requiring statistical knowledge. For example, when using decision trees to present demographic information on customers, the marketing department staff can read and interpret the graphical representation of the data without requiring statistical knowledge.

The data can also generate important insights on the probabilities, costs, and alternatives to various strategies formulated by the marketing department.

#### **2. Easy to prepare**

Compared to other decision techniques, decision trees take less effort for data preparation. However, users need to have ready information to create new variables with the power to predict the target variable. They can also create classifications of data without having to compute complex calculations. For complex situations, users can combine decision trees with other methods.

#### **3. Less data cleaning required**

Another advantage of decision trees is that there is less data cleaning required once the variables have been created. Cases of missing values and [outliers](https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm) have less significance on the decision tree’s data.

### **Disadvantages of Decision Trees**

#### **1. Unstable nature**

One of the limitations of decision trees is that they are largely unstable compared to other decision predictors. A small change in the data can result in a major change in the structure of the decision tree, which can convey a different result from what users will get in a normal event. The resulting change in the outcome can be managed by machine learning algorithms, such as [boosting](https://corporatefinanceinstitute.com/resources/knowledge/other/boosting/) and [bagging](https://corporatefinanceinstitute.com/resources/knowledge/other/bagging-bootstrap-aggregation/).

#### **2. Less effective in predicting the outcome of a continuous variable**

In addition, decision trees are less effective in making predictions when the main goal is to predict the outcome of a continuous variable. This is because decision trees tend to lose information when categorizing variables into multiple categories.

**Import Decision Tree regressor:-**

from sklearn.tree import DecisionTreeRegressor

# **LASSO:-**

Lasso regression is a type of [linear regression](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/) that uses [shrinkage](https://www.statisticshowto.com/shrinkage-estimator/). Shrinkage is where data values are shrunk towards a central point, like the [mean](https://www.statisticshowto.com/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 [muticollinearity](https://www.statisticshowto.com/multicollinearity/) or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

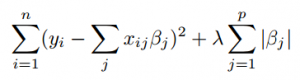
The acronym “LASSO” stands for Least Absolute Shrinkage and Selection Operator.

## **L1 Regularization**

Lasso regression performs L1 [regularization](https://www.statisticshowto.com/regularization/), which adds a penalty equal to the [absolute value](https://www.statisticshowto.com/integer/#abs) of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. [Ridge regression](https://www.statisticshowto.com/ridge-regression/)) *doesn’t* result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

## **Performing the Regression**

Lasso solutions are quadratic programming problems, which are best solved with software (like [Matlab](https://www.mathworks.com/help/stats/lasso.html)). The goal of the algorithm is to minimize:



Which is the same as minimizing the [sum of squares](https://www.statisticshowto.com/residual-sum-squares/) with constraint Σ |Bj≤ s (Σ = [summation notation](https://www.statisticshowto.com/calculus-definitions/summation-notation-sigma-function/)). Some of the βs are shrunk to exactly zero, resulting in a regression model that’s easier to interpret.

A [tuning parameter](https://www.statisticshowto.com/tuning-parameter/), λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

* When λ = 0, no parameters are eliminated. The estimate is equal to the one found with linear regression.
* As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when λ = ∞, *all* coefficients are eliminated).
* As λ increases, [bias](https://www.statisticshowto.com/what-is-bias/) increases.
* As λ decreases, [variance](https://www.statisticshowto.com/probability-and-statistics/variance/) increases.

If an [intercept](https://www.statisticshowto.com/problem-solving/find-intercepts-x-y/) is included in the model, it is usually left unchanged.

**Advantages and disadvantages of LASSO:-**

What are some of the main advantages and disadvantages of a LASSO regression model? Here are some of the main advantages and disadvantages of a LASSO model.

### **Advantages of LASSO regression**

* Automatic features selection. The main advantage of a LASSO regression model is that it has the ability to set the coefficients for features it does not consider interesting to zero. This means that the model does some automatic feature selection to decide which features should and should not be included on its own.
* Reduced overfitting. Another advantage of a LASSO regression is that the L1 penalty that is added to the model helps to prevent the model from overfitting. This makes intuitive sense because when the model sets feature coefficients to zero and effectively removes features from the model, model complexity decreases.

### **Disadvantages of LASSO regression**

* Biased coefficients. One of the main disadvantages of LASSO regression is that the coefficients that are produced by a LASSO model are biased. The L1 penalty that is added to the model artificially shrinks the coefficients closer to zero, or in some cases, all the way to zero. That means that the coefficients from a LASSO model do not represent the true magnitude of the relationship between the features and and the outcome, but rather a shrunken version of that magnitude.
* Difficult to estimate standard errors. Since the coefficient estimates in a LASSO model are biased, it is difficult to estimate accurate standard errors for them. This makes it difficult to do things like run statistical tests on them and build confidence intervals around them.
* Struggle with correlated features. Another downside of LASSO models is that they are very unstable when trained on data with correlated features. What usually happens is that one of the features gets selected somewhat arbitrarily and all of the other features that are highly correlated with that feature get effectively dropped from the model. This may lead someone to erroneously conclude that only the feature that was selected to remain in the model is important, when in reality some of the other features may be just as important or even more important.
* Generally unstable estimates. The estimates produced by LASSO models are known to be relatively unstable, which means that they can change a lot when trained on slightly different datasets. For example, if you bootstrap your data a few times to create a few different sample datasets, you might expect to see that different features get dropped from the model for each dataset. This can happen even if all of the datasets you are training on are very similar.
* Introduction of a hyperparameter. This disadvantage is more minor, but there is a hyperparameter that is introduced in LASSO models to regulate the size of the L1 penalty. That means that you have to go through hyperparameter tuning steps that you would not otherwise have to go through with a standard regression model.
* Other issues associated with standard regression models. LASSO regression models are also plagued by some of the same issues that affect standard regression models. Concerns surrounding interactions, outliers, and stringent model assumptions also apply to this family of models.

**Import LASSO**

from sklearn import linear\_model.Lasso

# 

# **RIDGE:-**

Ridge [regression](https://www.mygreatlearning.com/blog/what-is-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.

The cost function for ridge regression:

*Min(||Y – X(theta)||^2 + λ||theta||^2)*

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced.

* It shrinks the parameters. Therefore, it is used to prevent multicollinearity
* It reduces the model complexity by coefficient shrinkage
* Check out the free course on [regression analysis](https://www.mygreatlearning.com/academy/learn-for-free/courses/regression-analysis-with-excel-hands-on?gl-blog_id=20944).

## **Ridge Regression Models**

For any type of regression machine learning model, the usual regression equation forms the base which is written as:

*Y = XB + e*

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

Once we add the lambda function to this equation, the variance that is not evaluated by the general model is considered. After the data is ready and identified to be part of L2 regularization, there are steps that one can undertake.

## **Standardization**

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.

## **Bias and variance trade-off**

Bias and variance trade-off is generally complicated when it comes to building ridge regression models on an actual dataset. However, following the general trend which one needs to remember is:

1. The bias increases as λ increases.
2. The variance decreases as λ increases.

## **Assumptions of Ridge Regressions**

The assumptions of ridge regression are the same as that of linear regression: linearity, constant variance, and independence. However, as ridge regression does not provide confidence limits, the distribution of errors to be normal need not be assumed.

Now, let’s take an example of a linear regression problem and see how ridge regression if implemented, helps us to reduce the error.

We shall consider a data set on Food restaurants trying to find the best combination of food items to improve their sales in a particular region.

**Advantages:-**

* Avoids overfitting a model.
* They does not require unbiased estimators.
* They add just enough bias to make the estimates reasonably reliable approximations to true population values.
* They still perform well in cases of a large multivariate data with the number of predictors § larger than the number of observations (n).
* The ridge estimator is preferably good at improving the least-squares estimate when there is multicollinearity.

**Disadvantages:-**

* They include all the predictors in the final model.
* They are unable to perform feature selection.
* They shrink the coefficients towards zero.
* They trade the variance for bias.

**Import RIDGE**

from sklearn import linear\_model.Ridge

**8.**  **CONCLUSION**

**Conclusion from Model Training:-**

We saw that Sales column contains 172817 rows with 0 sale. So we created a new dataframe in which we removed 0 sales rows and tried to train our model. We used various algorithms and got accuracy score around 74%.

We werealso curious about the total dataset(including Sales = 0 rows). So we trained another model using various algorithms and we got accuracy near about 92% which is far better than previous model.

So we came to conclusion that removing sales=0 rows actually removes lot of information from dataset as it has 172817 rows which is quite large and therefore we decided not to remove those values.We got our best rmpse score from Random Forest model,we tried taking an optimum parameter so that our model doesnt overfit.

# **CONCLUSION FROM EDA:-**

1)From plot sales and competition Open Since Month shows sales go increasing from November and highest in month December.

2)From plot Sales and day of week, Sales highest on Monday and start declining from Tuesday to Saturday and on Sunday Sales almost near to Zero.

3)Plot between Promotion and Sales shows that promotion helps in increasing Sales.

4)Type of Store plays an important role in opening pattern of stores.

5)All Type ‘b’ stores never closed except for refurbishment or other reason.

6)All Type ‘b’ stores have comparatively higher sales and it mostly constant with peaks appears on weekends.

7)ssortment Level ‘b’ is only offered at Store Type ‘b’.

8)We can observe that most of the stores remain closed during State Holidays. But it is interesting to note that the number of stores opened during School Holidays were more than that were opened during State Holidays.

**References-**

1. Geeks for Geeks
2. Analytics Vidhya
3. Nvidia
4. scikit-learn