**Sales Prediction using Machine Learning**

So here are few steps guidance on how to predict sales using machine learning



Introduction:

So here you will get a complete analysis of a Sales Prediction model. Here I will cover all major insights that we used to be see in our ML model or projects, we will do the complete analysis from data visualization to model building and finding

the key observations from the analysis, that will help us to predict the results.

**So, following are the required steps:**

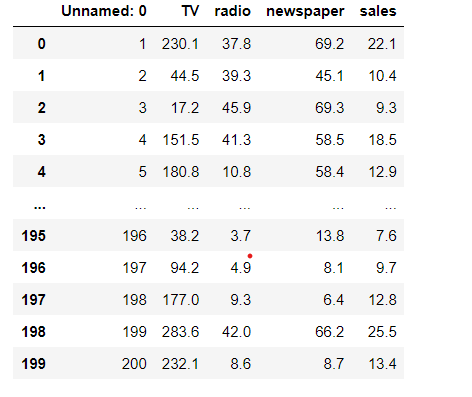
**1.Problem definition**

* So, the problem is that when a company enters a market the distribution strategy and channels it uses are keys to its success in the market
* Because an effective distribution strategy under efficient supply- chain management open doors for attaining competitive advantage and strong brand equity in the market it is a component of the marketing mix that I cannot be ignored.
* The distribution strategy and the channel design have to be right the first time.
* This case study of sales channels say includes the detailed study of TV, radio and newspaper channel.
* Thus predicts the total sales generated from all the sales channel.
* So here we have clarified what is the problem.

**2. Data Analysis**

Now we are just analyzing our sales dataset, and let’s see what

insights we gain from it. So, I am sharing a picture of the dataset, I have used for the analysis.



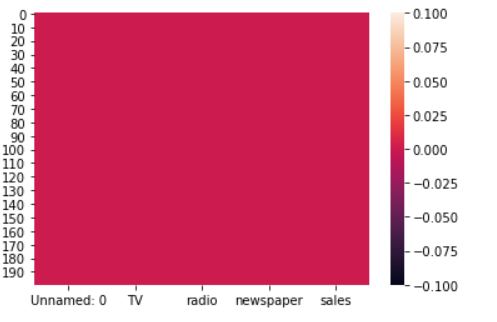
* We can see that we have 200 rows and 5 columns in the dataset.
* We have 5 columns Unnamed :0, TV, radio, newspaper and sales.
* We check the properties like shape, unique, dtypes etc.
* We have sales column as our target and is continuous in nature, thus it is a Linear Regression problem.
* Now I will do the further analysis according to our problem type, which is linear regression is a type of Supervised learning.

**3.EDA Concluding Remarks**

Now we will see the complete EDA of the dataset. By doing following steps

1. **Check missing values**
2. **Statistical Summary**
3. **Univariate Analysis**
4. **Bivariate Analysis**
5. **Multivariate Analysis**
6. **To check Skewness**

Let’s discuss the above steps in detail:

1. We can see whether null values are present in the dataset or not, we also use heatmap like I have done

We can see that no null values are present in the dataset.

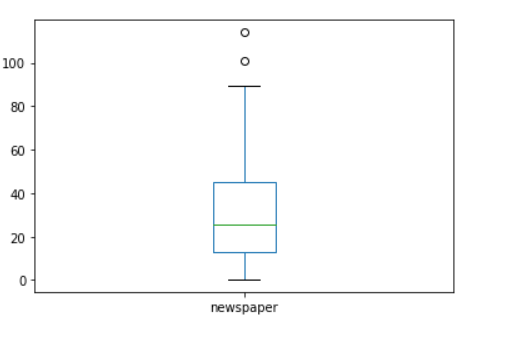
**2)** Statistical summary gives information about the mean, median, std, min, max etc.

key observations:

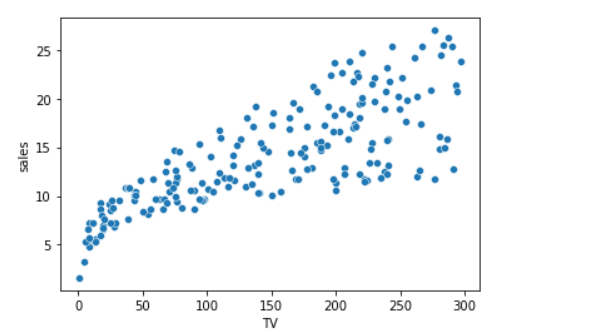
1 The difference in mean and median is almost similar.

2 There is small difference in 75% percentile and max in columns named Unnamed: 0, TV and radio which shows that no outliers are present in it.

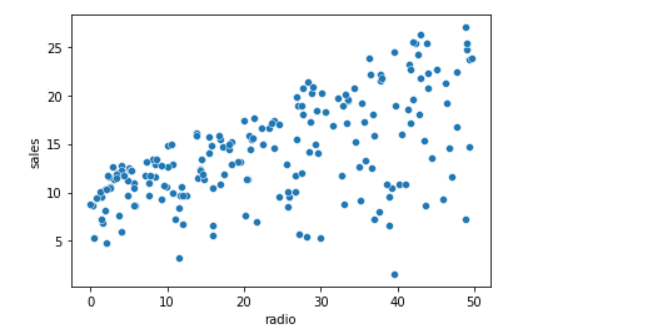
3 we can see the difference in 75% percentile and max in newspaper column which shows that few outliers are present in it.

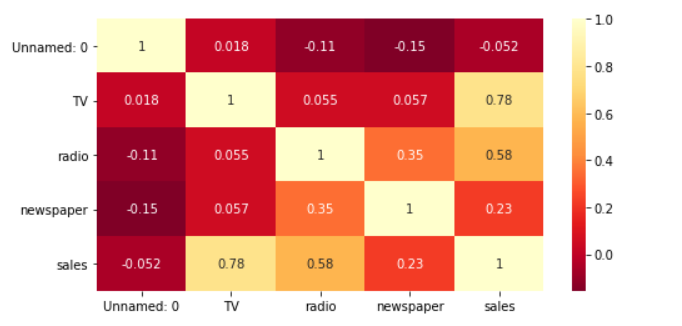
**3)** In univariate analysis I have plotted boxplots, from these plots we can see the mean, median, max, min and we can also see whether outliers are present or not, as we can see in below picture, we can see that outliers are present in newspaper column.

**4)** Now in Bivariate analysis I have used Scatter plot to see the relation of each column with the sales column.

we can see the scatter plot of TV column with the sales. Here we can see the positive relation between the sales and TV, as the TV advertisement increases sales also increases.

We can see also see the plot of radio column also, here we can see that the sales increases with the increase in radio advertisement but sometimes sale is decreased too.



**5)** In multivariate analysis we can check the correlation matrix of the dataset, which clearly shows us the relation of each column with other column, here is the correlation matrix.

Observations:

Light shades are highly correlated.

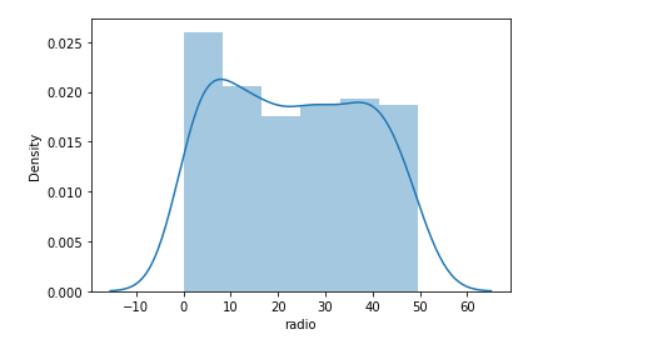
Sales are highly positively correlated with Tv column.

Sales are highly negatively correlated with Unnamed:0 column.

Sales are slightly positively correlated with radio column.

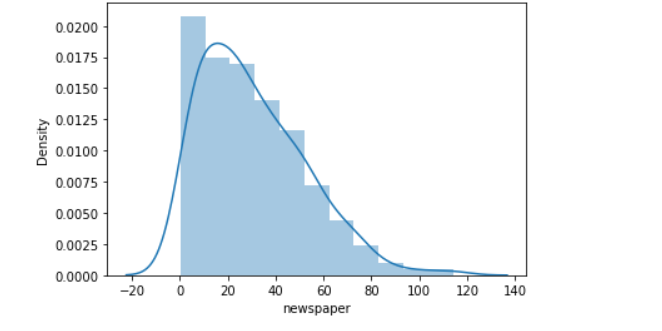
Sales are negatively correlated with newspaper column.

**6)** Now I have checked the skewness of each column by plotting density plot, to check whether the curve is normally distributed or not, how much skewness is present? We have skewness in many columns let’s see in radio column



Here we can see that curve is not normally distributed, thus we can see the skewness clearly.

We can also see the skewness of newspaper column in the below picture.



Here we can also see that the curve is slightly skewed from the right.

**4.Pre-processing pipeline**

we have many steps included in pre- processing like Data cleaning, Data reduction, Data integration etc. Let's discuss each of them in detail the steps we have done in our project

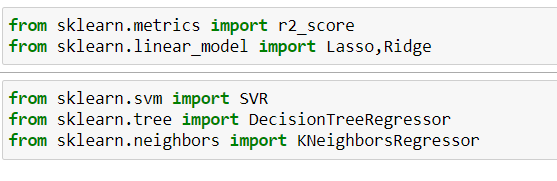
according to our requirement.

1. We have drooped the rows which are negatively correlated, here we have dropped the Unnamed: 0 column from the dataset due to its highly negative correlation with the sales column. This column impacts our data negatively, thus we dropped that column.
2. Now we are removing the outliers present in our data, we have two methods to remove outliers one is zscore and other is by using IQR method here I have used zscore method to remove all the outliers present in our dataset, I have very less outliers in my dataset thus only two rows are deleted from the dataset as outliers.
3. Here we have seen above that no null values are present in our dataset, so here is no need to handle missing values, if there were any missing data then we have to treat it with suitable method, but here no need.
4. Next is the we check whether any column is present in string format or not, if any column is present in string format, then we have to change it in integer format by applying Encoding technique, we have two methods in encoding one is One hot encoding and other is Label Encoding, but here in our dataset we have no need to apply encoding technique because all the columns are already present in integer format. So, let’s move on to the next step.
5. So now we are removing skewness from our dataset, as we have seen that skewness is present in all the columns except the sales column, to remove the skewness we have separated the target variable and the independent variable from the dataset. As we know that skewness between –0.5 to 0.5 is acceptable but more than it is not, so skewness except this range should be treated by using suitable method. So, here I have used cube Root method to handle negative skewness and square root method to handle positive skewness, till now we have treated the skewness by suitable method and removed skewness from our dataset.
6. Next point which comes is the feature engineering but here in this dataset we do not need it, because we don’t have special characters, etc. So here we have no need to apply Feature Engineering at all.
7. The last thing which I can see is the **standardization** technique, we use this technique to scale our data. We have two methods to scale our data first one is standard scaler and the second one is min max scaler. We use these techniques only when there is huge difference between the ranges of any 2 columns, that’s why we use scaling. **Standard scaler** is used when data is normally distributed, its changes the mean=0, std=1 and the value ranges between –3 to +3. **Min –Max scaler** is used when data is not normally distributed this method is also called normalization, it changes the data with mean=0, std=1 but range is 0 to 1. In our dataset we have seen there is no huge difference in ranges of the columns, thus here we have no need to apply standardization technique on our dataset.
8. The last which I can see is the PCA technique, this technique is used only when we have, we large numbers of columns and it's difficult to manage them all, but here in our dataset we have only 5 columns thus here we have no need to use this technique.

**5.Building Machine Learning Models**

Now we build a machine learning model, we will use multiple algorithms, as we know we are working on a regression problem so here we will only use regression models like linear

Regression, SVR, DecisionTreeRegressor, KNeighborsRegressor, Lasso Regression, Ridge Regression. To use all these, we have to import each model from scikit learn as follows:



Firstly, we have to create train test split and thus we use train data for training our model and test data for testing our model performance. Here I have splited the data in 70% as train and 30% as test data. I have finded the best random state using the regression model which is 52, I used this random state to train all the models. Now we use train data for training our model and test data for testing our model performance.

Since we are using regression model's thus, we different evaluation matrix like mean absolute error, mean squared error and Root mean squared error for all the models

The error in the models shows the performance of the model if error is least the model is performing well, but if the error is more model is not performing good.

For our dataset we have checked the error and founded that the least error is coming from the Decision Tree Regressor.

So now let’s use evaluation matrix r2\_score to see the score for all the models, after checking the r2\_score I have finded that I am getting maximum r2\_score with DecisionTreeRegressor is 95%.

The picture which comes around us is the **cross-validation** **technique,** as we know the score is also due to overfitting, thus we use cross validation method to come over it.

**Cross**-**Validation** is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to **validate** the model. ... The basic form of **cross**-**validation** is k-fold **cross**-**validation**.

Here in my model, when I checked the cross validation, I use cv=5 folds, the best cross-validation score is coming out to be for Decision tree regressor.

Now I have used hyper parameter tunning to find the best parameter for our model by using GridsearchCV.

I have applied gridsearchcv on all the models and finded the best parameter for all the models and used these parameters in our model, now I am getting the best accuracy with decision tree regressor with(parameter=mse) .

Now we can conclude that decision tree regressor is the best model for our dataset, as we have seen we are getting best r2\_score with that model, least error, least difference in r2\_score and cross validation score, so it is our best model for this project.

So last step is to save our model to use it in future for predictions, we have two techniques for saving the model. First is using joblib and the second is by using pickle.

I have saved my model DTR using joblib, so that I can use it in future and predict the sales using this model for future use.

The last thing we have to do is concluding remarks:

**6)Concluding Remarks**

Let 'see the conclusion in below points:

* The main goal of our project is to solve the problem and predict the sales.
* For this we used machine learning skills and solved the issue.
* We have done the complete analysis of the data using EDA, univariate, bivariate, multivariate, checking correlation, checking skewness, checking for outliers, checking for missing values by doing all this analysis we have collected the information about the data, whether it is skewed, having missing values or not etc.
* Next, I have done the pre-processing of the data and solved all the issues that we finded during EDA like Outliers, skewness etc.
* The last is the model building I used regression algorithm and different evaluation matrix to prepare the models and finded decision tree regressor as my best model
* And at last, we can make predictions for sales channel using our model.

Author Name=Neha