PROBLEM STATEMENT

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- We have a dataset named "Leads.csv" which contains the details of customers who are both converted and not converted which is depicted in the columns "Converted" along with other details like their identification number , preferances , page visit frequencies,City , How they came to know about the course

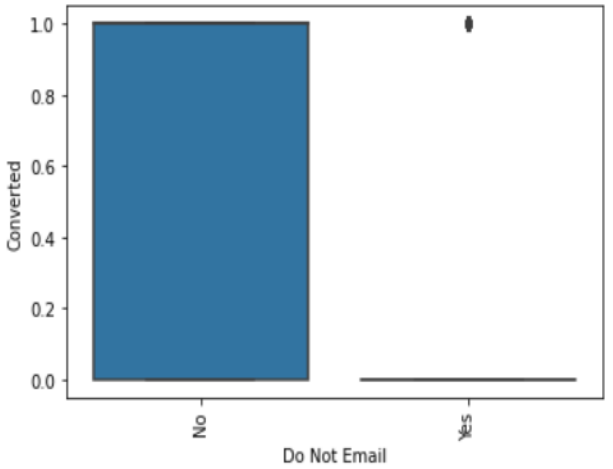
- For any kind of dataset we receive there are certain kind of basic details which we must know before starting any model building ,such as their shape ,column names ,etc…

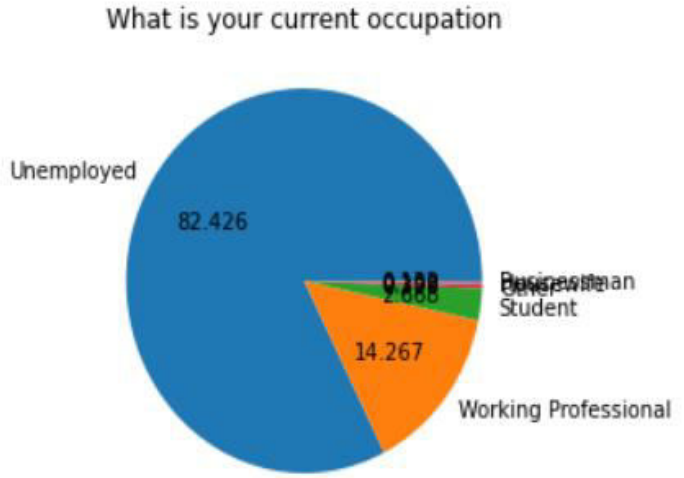
- After doing the exploratory data analysis for our dataset "Leads.csv" we could see that we have a dataset with 9240 entries and 37 columns

- The dataset contains both continuous data such as Totalvisits,pageviews,profilescore ,activity score etc and categorical data such as DonotEmail,LastActivity, Country etc..

- After dropping all the columns which have more than 40% null values and rows which contain null values our final dataset has 4535 rows and 12 columns.

- From the univariate analysis ,we could see that:
  - Majority of the Converted users has opted for getting email communications
  - Almost equal number of people from all Specializations has been Converted
  - Majority of the users came to know about the course from "Direct Traffic and Google"
  - Last Activity of majority of users is "Email Opened and SMS Sent"
  - 82% of users are Unemployed

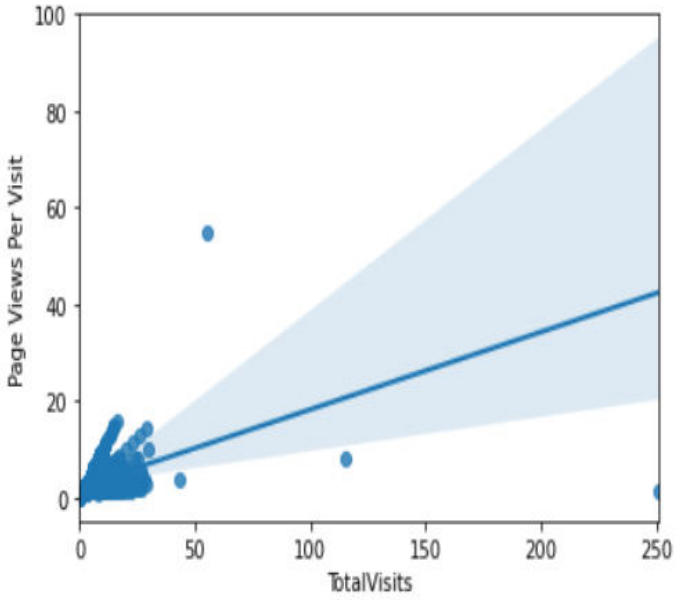


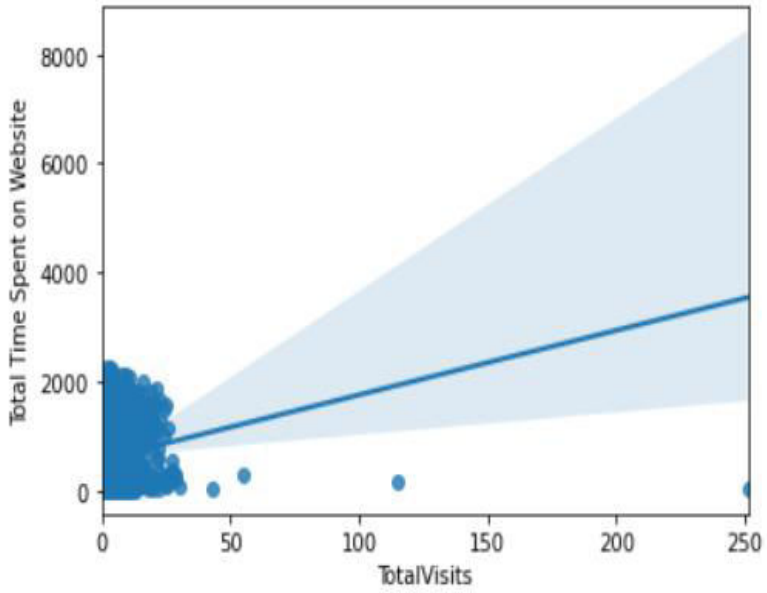


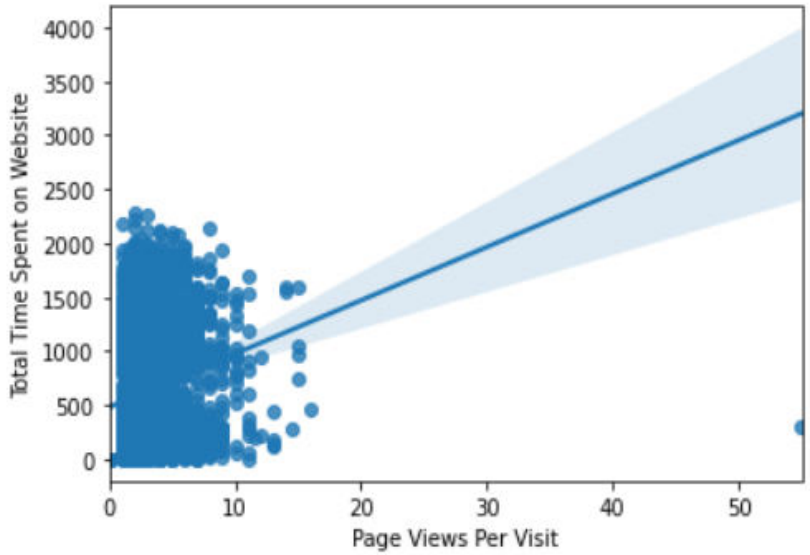
From Bivariate analysis of three variables "TotalVisits","Total Time Spent on Website","Page Views Per Visit" we could infer that :

- TotalVisits and Page Views Per Visit have a linear relationship

- Although not that clear relationship (Total Time Spent on Website & Page Views Per Visit) and (Total Time Spent on Website & TotalVisits) also has a similar linear relation

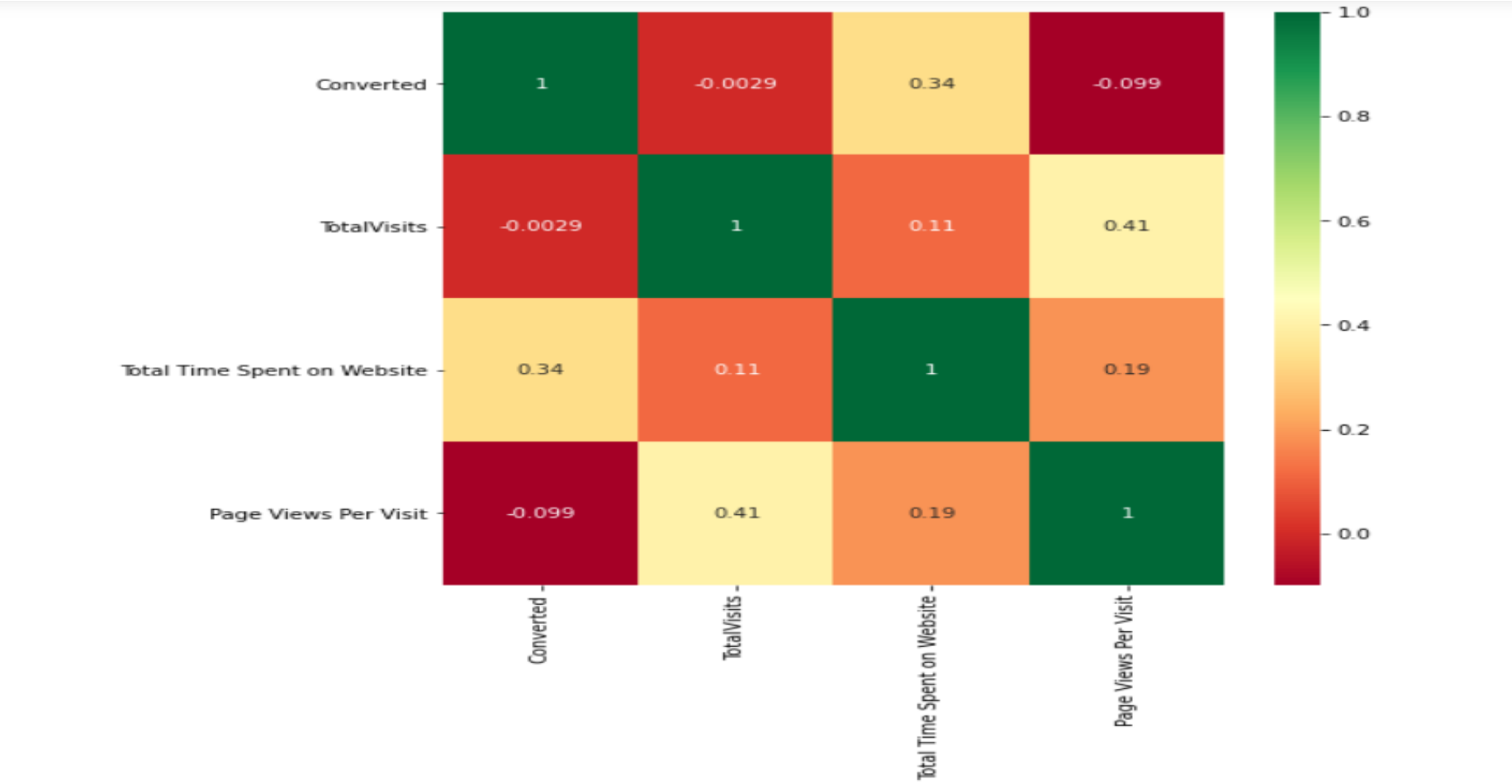






HeatMap

- We can infer from the heatmap that
  - Page Views Per Visit and Converted has a negative covariance
  - Highest covariance is between Page Views Per Visit and TotalVisits which is 0.41



Data Preparation before model building

- Columns "Do Not Email" and " A Free Copy of Mastering the Interview " which are having values as "yes" and "no" are converted to "0" and "1".

- All the categorical variables are converted to dummy variables.

- After adding the new columns with dummy variables and dropping old categorical columns we have a final dataset with 4535 rows and 71 columns

Correlation Check

- After checking the correlation among different columns ,w e could see the correlation for following columns are more than 0.8:
  - Lead Origin_Lead Add Form & Lead Source_Reference = 0.967373
  - Last Activity_SMS Sent & Last Notable Activity_SMS Sent = 0.892981
  - Last Notable Activity_Unsubscribed & Last Activity_Unsubscribed = 0.886936
  -
  - Last Notable Activity_Email Opened & Last Activity_Email Opened = 0.875955
- All these are highly correlated columns ,so dropped one from each of the above pairs
- Dropped "Lead Origin_Lead Add Form ,Last Notable Activity_SMS Sent , Last Notable Activity_Unsubscribed , Last Notable Activity_Email Opened"
- The final dataset has 4535 rows and 67 columns

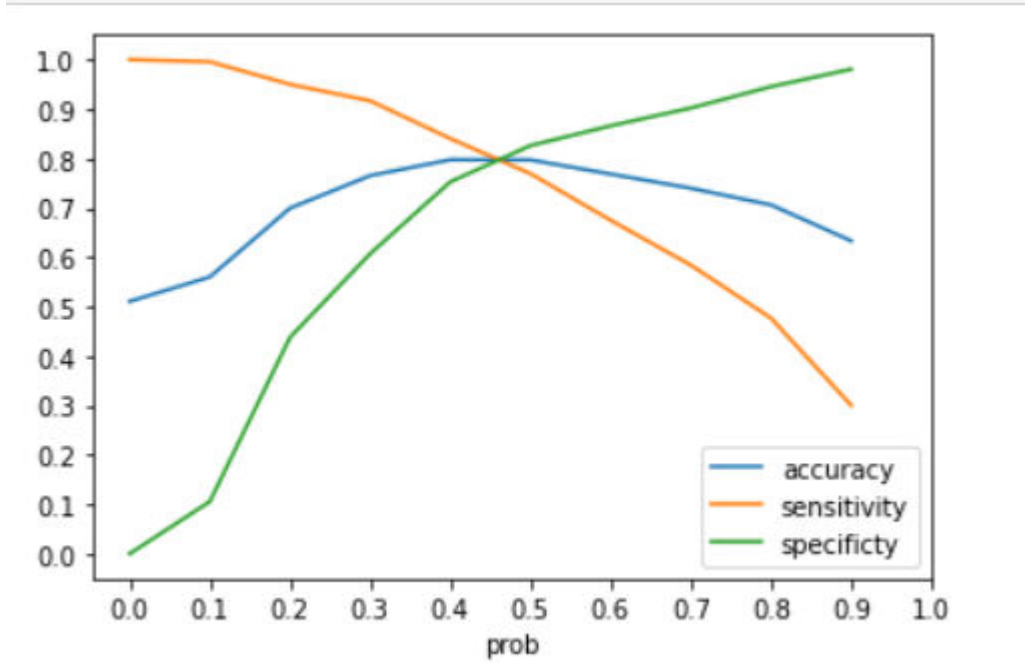
Test Train split and Feature Scaling

- 80% of data has been used for training and 20% for testing
- MinMax Scaling has been applied for feature scaling.

TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	S
0.055777	0.119159	0.127273	1	0	0	0	0	0	1	
0.019920	0.005607	0.030364	1	0	0	0	1	0	0	
0.015936	0.014019	0.072727	1	0	1	0	0	0	0	
0.023904	0.073832	0.054545	1	0	1	0	0	0	0	
0.043825	0.088785	0.066727	1	0	0	0	1	0	0	

MODEL BUILDING

- We have selected 15 columns from 67 columns using RFE method
- Using Generalized Linear Model Regression and VIF to check the p value and multicollinearity of the variables
- After dropping the three columns – "Lead Source_Welingak Website , Last Notable Activity_Had a Phone Conversation , What is your current occupation_Housewife" which had a high p value of 0.999 which is higher than the threshold 0.05
- We have arrived at Model 4 which has VIF and p values for all variables under threshold.
- After model evaluation , we have finalized the cutoff value as 0.47



Using cutoff as 0.47 for our train model we have received the following values :

- Accuracy : 0.800252047889099
- Specificity : 0.8074694140373471
- Sensitvity : 0.7933374460209747
- Precision : 0.8113564668769716
- Recall : 0.7933374460209747

	Converted	Converted_Prob	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Final Predicted	Lead_Score
0	0	0.185151	1	1	0	0	0	0	0	0	0	0	0	18.52
1	0	0.135024	1	1	0	0	0	0	0	0	0	0	0	13.50
2	1	0.247204	1	1	1	0	0	0	0	0	0	0	0	24.72
3	0	0.334283	1	1	1	1	0	0	0	0	0	0	0	33.43
4	0	0.192593	1	1	0	0	0	0	0	0	0	0	0	19.26

After applying the model on our test set we have the following values :

- Accuracy : 0.7876561351947098
- Specificity : 0.7882882882882883
- Sensitvity : 0.7870503597122303
- Precision : 0.7950581395348837
- Recall : 0.7870503597122303

	Converted	Converted_Prob	Final Predicted	Lead_Score
0	0	0.266142	0	26.61
1	0	0.202790	0	20.28
2	1	0.883448	1	88.34
3	1	0.981265	1	98.13
4	0	0.163276	0	16.33

So from our final model we could understand that :

From a business point of view we can concentrate more on these features of the customer data like TotalVisits , Total Time Spent on Website both of which has the highest positive coefficients ,so will have a positive impact on the result and Page Views Per Visit which has highest negative impact .