Lead Scoring ASSIGNMENT

PRESENTED BY

ABHINAV SINGH

CONTENTS

PROBLEM STATEMENT

The agenda is to build a logistic Regression Model that predicts the score of the prospects indicating their probability of getting converted as a customer .

WORKFLOW

1. Importing Libraries And Warnings
2. Reading The Data Files
3. Missing Values And Treatment
4. Checking And Validating Data Types
5. Univariate Analysis
6. Data Preparation
7. Train-Test Split
8. Model Building
9. Model Evaluation on Training Data
10. Final Model Evaluation on Test Data
11. Finding out the leads to be contacted
12. Finding the best Parameters for the Model

IMPORTING LIBRARIES

- Imported the necessary libraries for the EDA and Model Development Like Pandas,Matplotlib , Seaborn and Numpy
  - import numpy as np
  - import pandas as pd
  - import matplotlib.pyplot as plt
  - import seaborn as sns
- Imported Warnings to filter the warnings
  - import warnings
  - warnings.filterwarnings('ignore')

Loading DataSet

- The “Score” variable is our Leads.csv DATA SET indicates the probability of a prospect to get converted as customers.
- The leads.csv data set has 36 independent variables and 1 “Score “ as dependent variable'
- The leads.csv contains 9240 records of the prospects .

ROUTINE CHECKS

Routine checks on both the application.csv and previous application.csv were done.

Few of the routine checks done were

- Getting the information of data type
- Getting the Shape and Size
- Getting the statistic summary

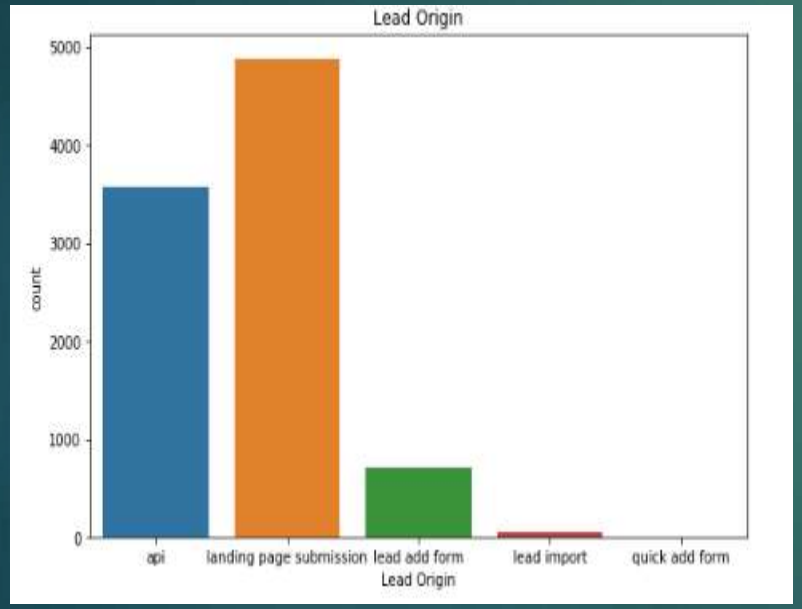
All these checks are highlighted in the notebook

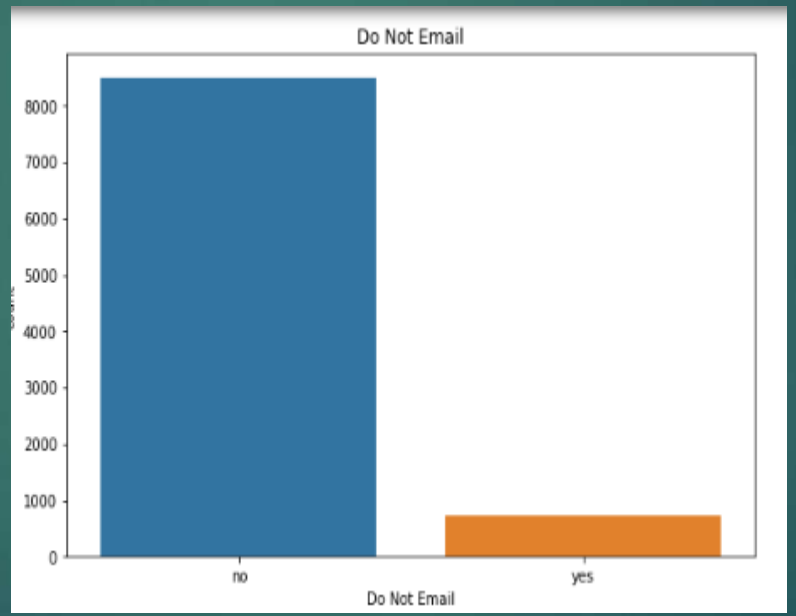
Missing Value Treatment

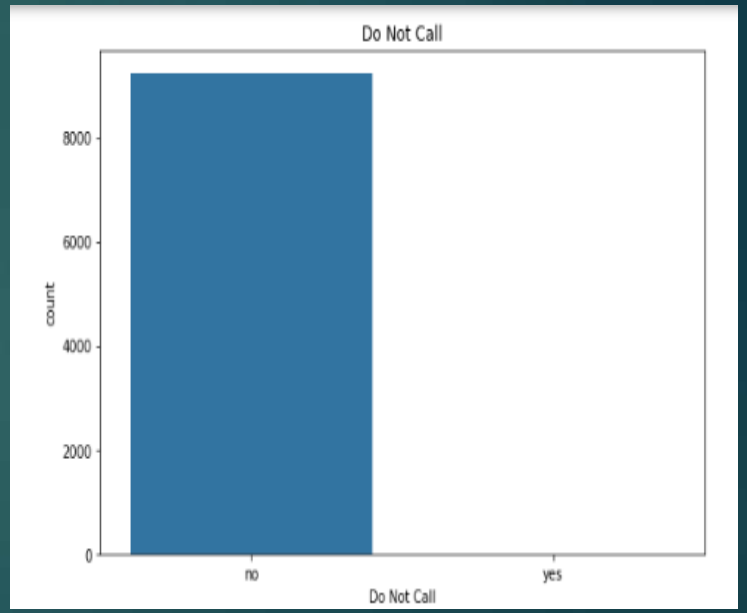
- In the dataset all the columns with more than 45% of missing values have been dropped.
- All the columns that had only one value were dropped as model won't be able to learn anything from these and its complexity would increase.
- For categorical columns the where select was there it was replaced with 'NAN' and imputed with 'Not provided'.

UNIVARIATE ANALYSIS

We can see that the leads origin with most counts is “landing page submission” followed by “api. Also most of the prospects do not prefer to be mailed or called.

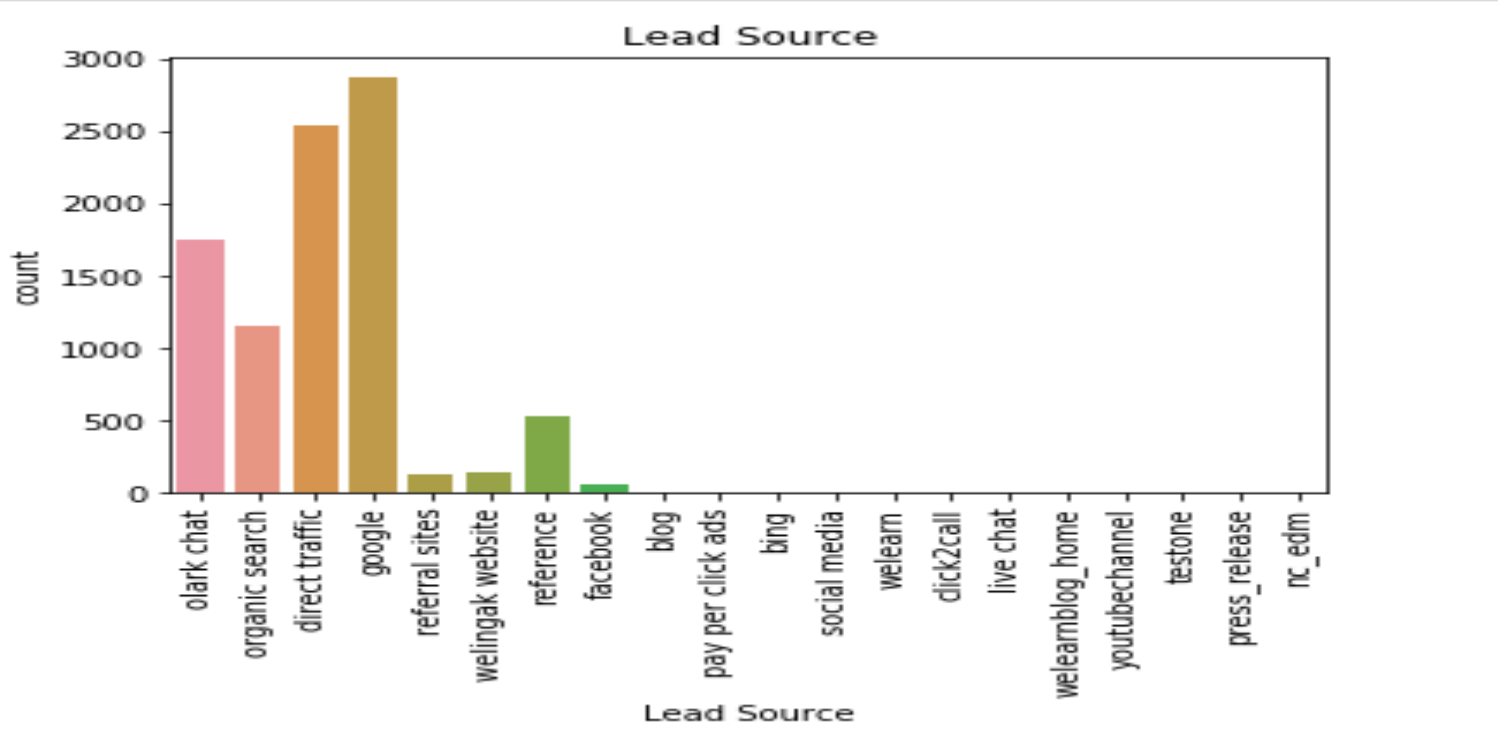






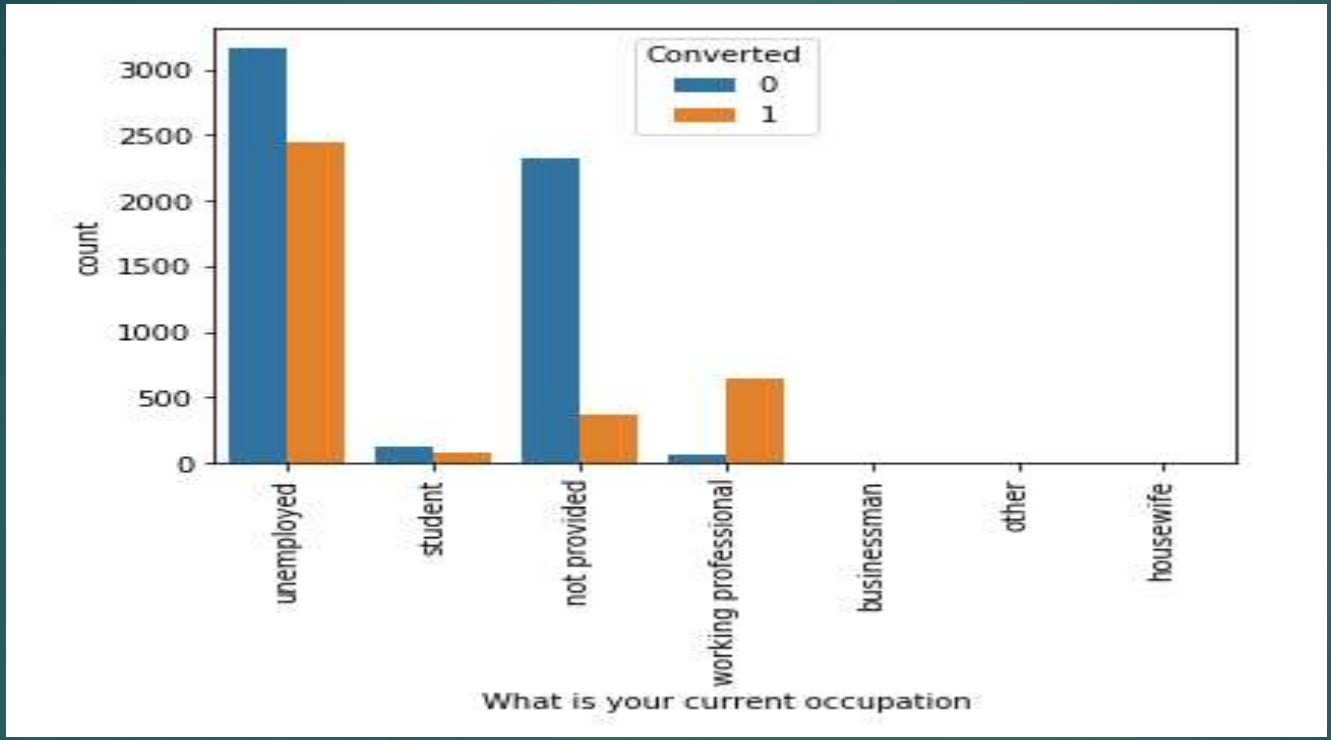
UNIVARIATE ANALYSIS

Google is the highest Lead source among all the lead sources as most people have presence over google and it is followed by direct traffic.



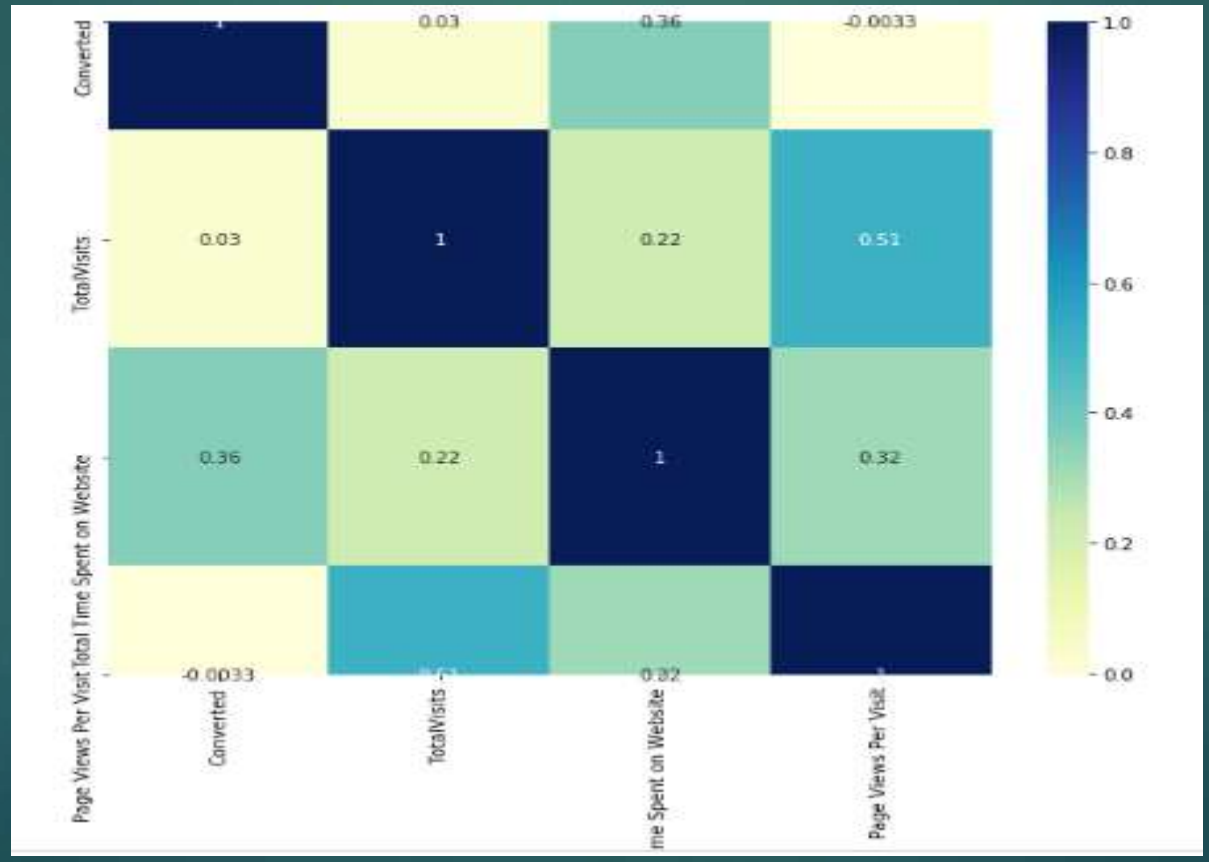
UNIVARIATE ANALYSIS

We can see most of the converted leads come from people currently unemployed and second place working professionals



Correlation Checks

There are no high correlation values, the highs one is the total Visits vs page views per visit.



Data Preparation

- Data preparation is an important step in model building and we started by converting the categorical columns with two values into numerical columns with 0 and 1 values.
- The next step for the remaining multivariate categorical columns was to create dummies as logistic regression model won't support categorical features in the data set.
- After creating the dummy variables the original variables were dropped from the dataset .
- Now we had 149 features after creating the dummy variables.

Train Test Split

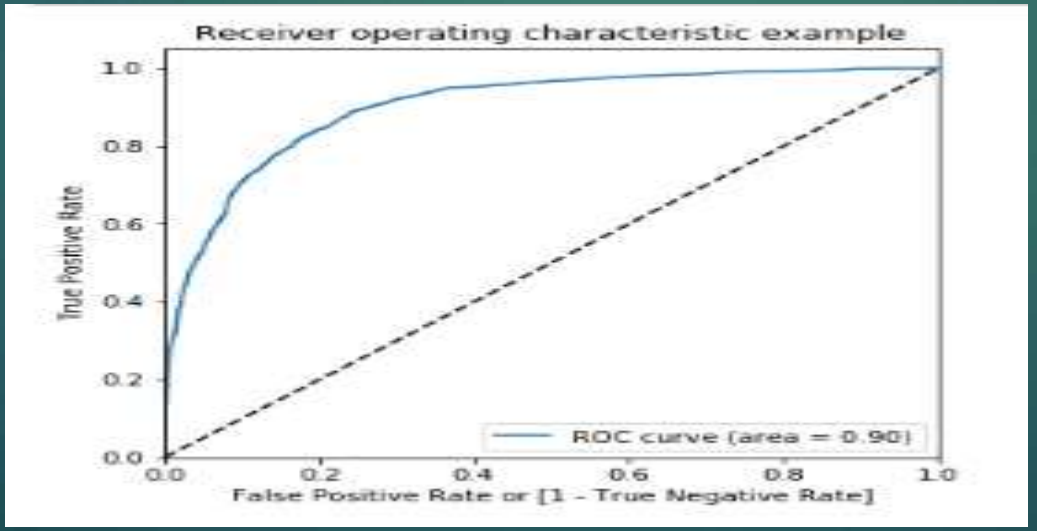
- After Preparing the data we did split the data into tranining and testing dataset for model building and evaluation.
- The target variable “score” was separated from this dataset before splitting.
- The dataset was split in 70:30 ratio that is 70% of the original data was used as training data set and the 30% of the original data was used as a testing data set.
- Some numerical features were scaled using minmax scaler so that all the features are in same kind of scale.

Model Building

- Initially for model building we decided to use RFE to automatically select the most important 20 variables .
- In the second step we created stats model for more detailed results using the above 20 variables and checked them using P values.
- In subsequent models the variables were removed basis high P values.
- Once all the Variables had P values in range we checked for VIF and removed variables on that basis.
- The final model had 16 variables as the main predicting variables . The list of these variable's can be seen in the python notebook.

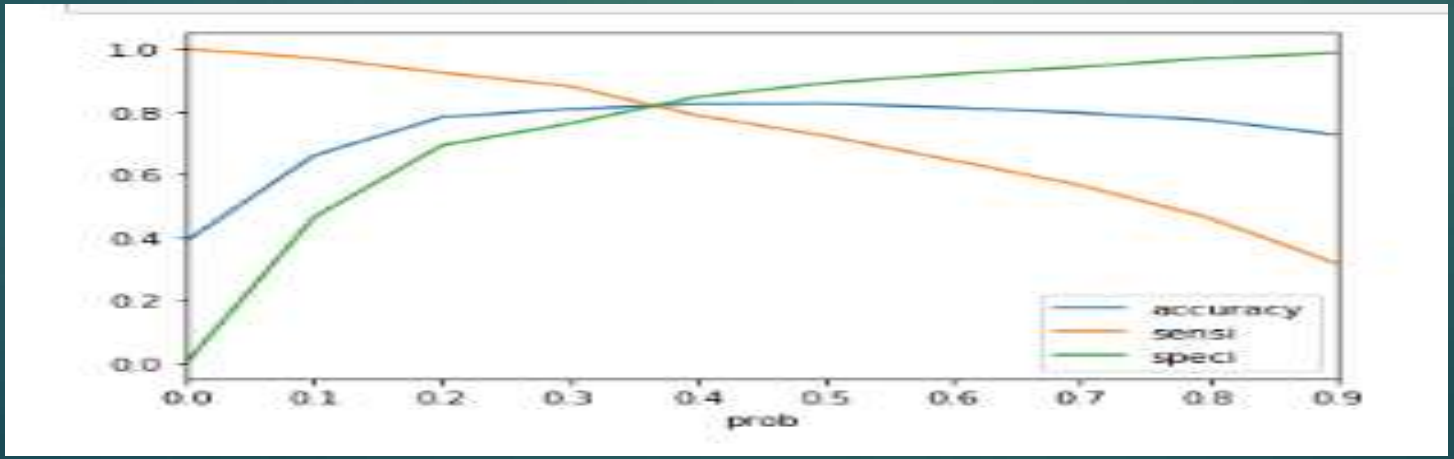
Model Evaluation on Training Data

- Once we got the final features predictions were made on the training data to check the accuracy of the model during training period.
- The overall Accuracy of the model on training dataset came to be 82.5%.
- Other metrics like Sensitivity , Specificity and ROC curve were also measured to check the model performance
- Sensitivity of the model was 72.31% while specificity was 89%.
- The AuROC was 90%.



Finding Optimal Cutoff and Evaluation of Model

- At start for default we took 0.5 as cutoff for predicting 0 or 1 but we needed to find accurate cutoff for our model to be reliable.
- For this the sensitivity , specificity and accuracy for different threshold were checked and can be seen in the python notebook .
- After plotting the graph for all the above metrics at different threshold the best cutoff seemed to be 0.35.



Finding Optimal Cutoff and Evaluation of Model

- After choosing the cutoff to be 0.35 the model evolution on different metrics was done again.
- The accuracy of the final model with 0.35 as cutoff came to be 82% approx.
- Sensitivity was 83%
- Specificity was 81 %
- Precision was 80%
- Recall was 72%
- These all metrics indicated our model is performing well on training but we still need to evaluate it on test data set to be sure if our model is stable and reliable .

Model Evaluation On Test Dataset

- The model performance for the test data set was checked to see how the model performs on unseen dataset.
- The test dataset was given similar treatment as of training data set before making any predictions
- The Accuracy of the model on test data set was 83% which is close to training dataset and hence it signifies our model is stable.
- Sensitivity was 83%
- Specificity was 82%
- All the above metrics evaluation on test and train data were satisfactory and indicated the model was performing well.

Finding the Leads to be contacted

- Basis the model prediction we thought a cutoff of lead score greater than 80 would be best for contacting the prospects.
- We suggest any prospect having lead score of 80 and more can be converted to a customer.
- Total 520 prospects had a lead score of greater than 80 and should be contacted first.

	Prospect ID	Converted	Converted_prob	final_predicted	Lead_Score
12	8610	0	0.834668	1	83
13	4776	1	0.998029	1	100
18	1618	0	0.990388	1	99
20	1869	1	0.909839	1	91
22	7215	1	0.989193	1	99
...	...	...	...	...	...
2706	6021	1	0.873060	1	87
2714	4227	1	0.862581	1	86
2720	8090	1	0.990944	1	99
2722	2497	1	0.997007	1	100
2724	7716	1	0.832271	1	83

520 rows × 5 columns

Best Parameters for the Model

- The 16 best parameters along with their importance can be seen below.

```
Lead Profile_lateral student                                    23.201148
TotalVisits                                                      5.452723
Total Time Spent on Website                                      4.491025
Lead Source_welingak website                                     3.087109
What is your current occupation_working professional             2.349501
Lead Origin_lead add form                                        1.949665
Last Activity_had a phone conversation                           1.817455
Lead Profile_potential lead                                      1.532205
Last Activity_unsubscribed                                       1.470577
Country_not provided                                             1.433053
Last Activity_sms sent                                           1.197530
Last Notable Activity_modified                                  -0.797754
Last Activity_olark chat conversation                           -0.869864
What matters most to you in choosing a course_not provided      -0.882610
Do Not Email_yes                                                -1.457568
const                                                           -2.234499
Lead Profile_student of someschool                              -2.393066
dtype: float64
```

Recommendations

- The company should make more calls where lead profile is lateral students.
- The company should make calls to the leads who are working professionals
- The company should contact the leads who spent more time on websites
- The company should make more calls where last activity was had conversation.
- The company should make more calls whose last activity was sms sent.
- The company should not make calls where last activity was do not sent mail yes.

THANK YOU