LEAD SCORING CASE STUDY

By,

KEERTHANA DS

PROBLEM STATEMENT

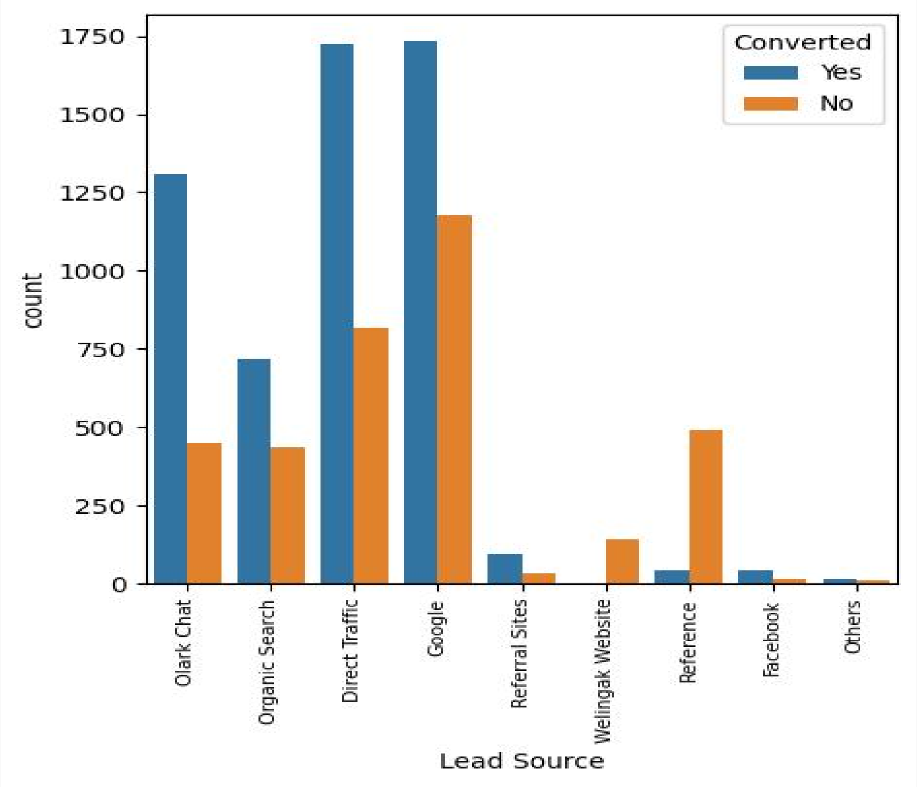
- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- When these people fill up a form providing their email address or phone number in the website, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. There are a lot of leads generated in the initial stage but only a few of them come out as paying customers.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

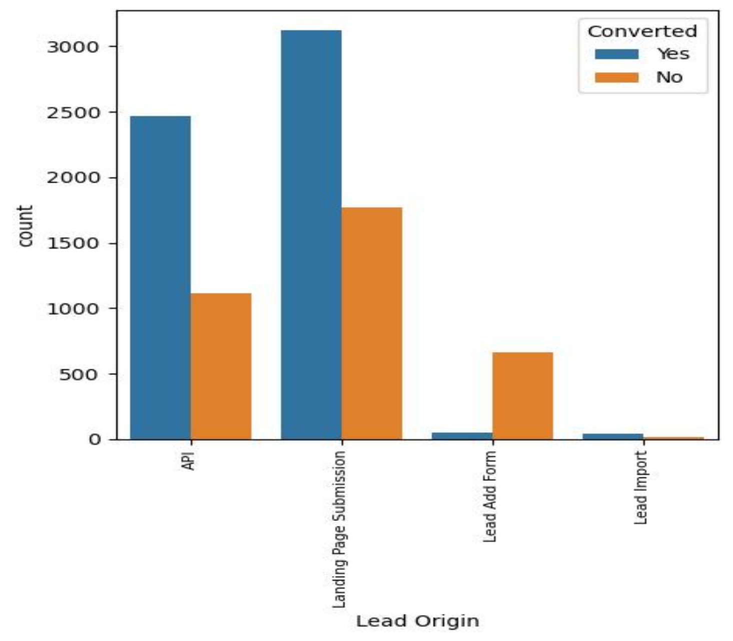
EDA STEPS

- Reading and Understanding the data
- Data Preparation
  - Missing value treatment
  - Handling redundant variables
  - Outlier treatment
  - Data Analysis of the categorical and numerical variables
  - Converting binary variables
  - Creating dummy variables
- Train-Test Split
- Feature Scaling
- Model Building
- Plotting ROC
- Finding optimal threshold
- Predictions on test set
- Model evaluation

EDA

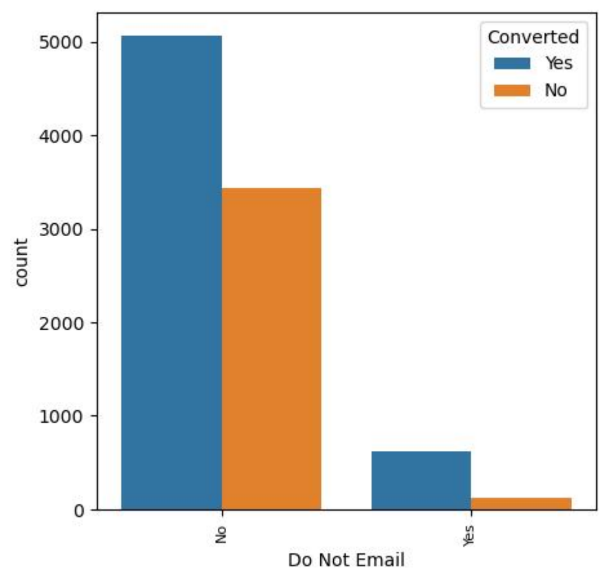
Categorical Variable Analysis

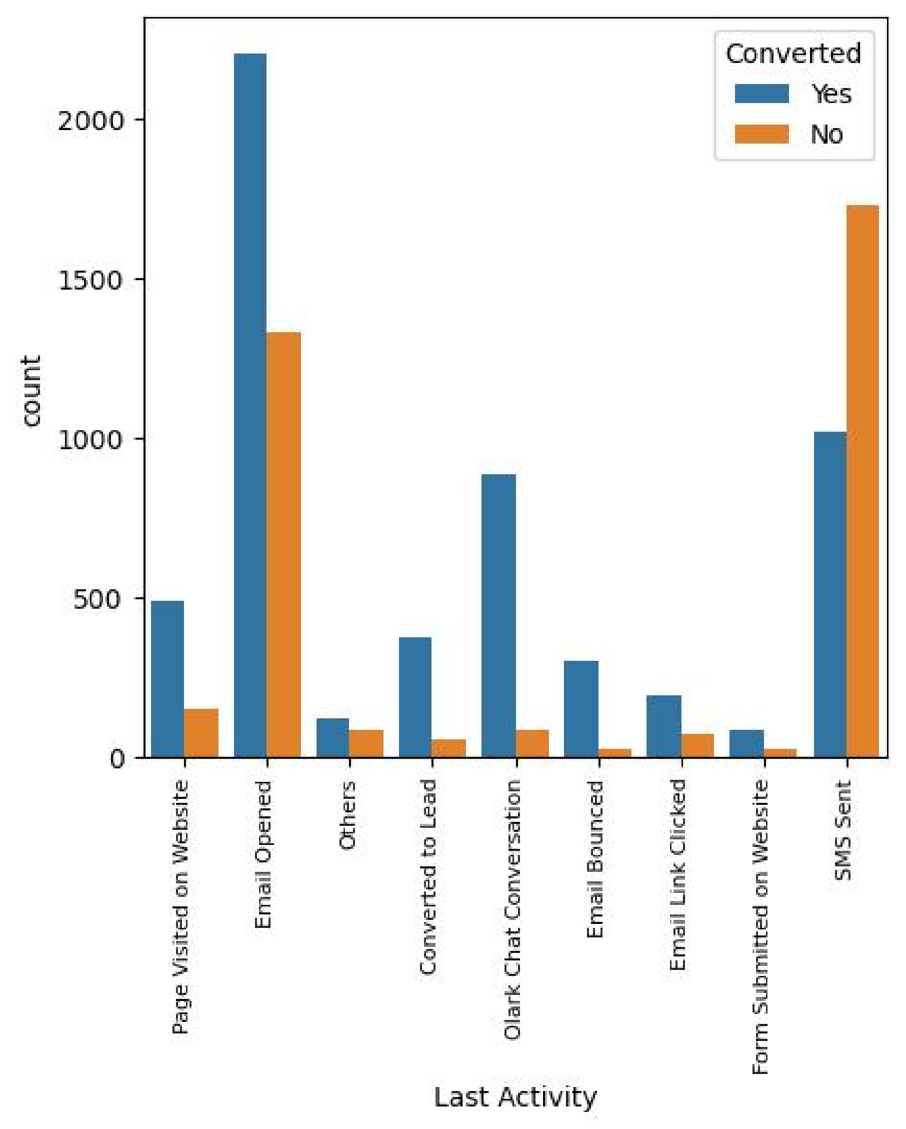




EDA

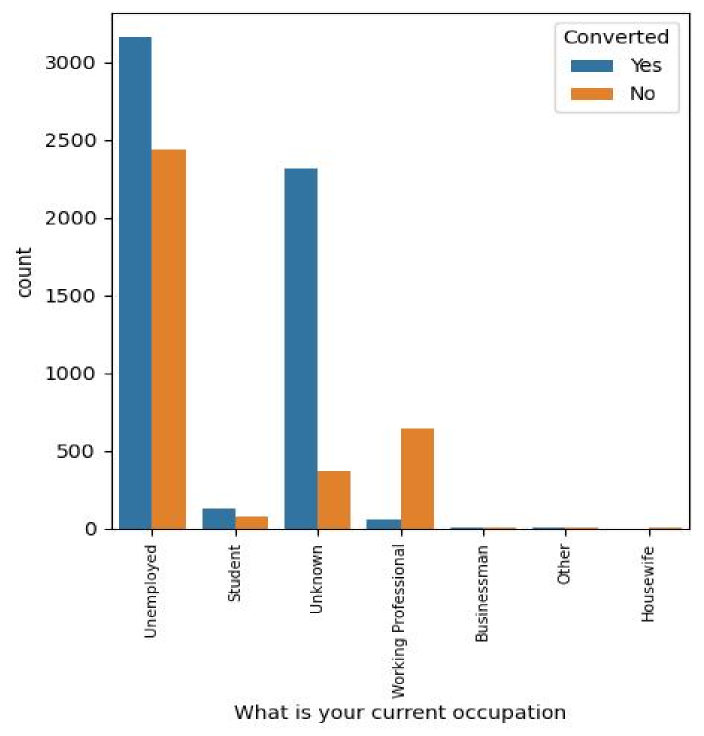
Categorical Variable Analysis

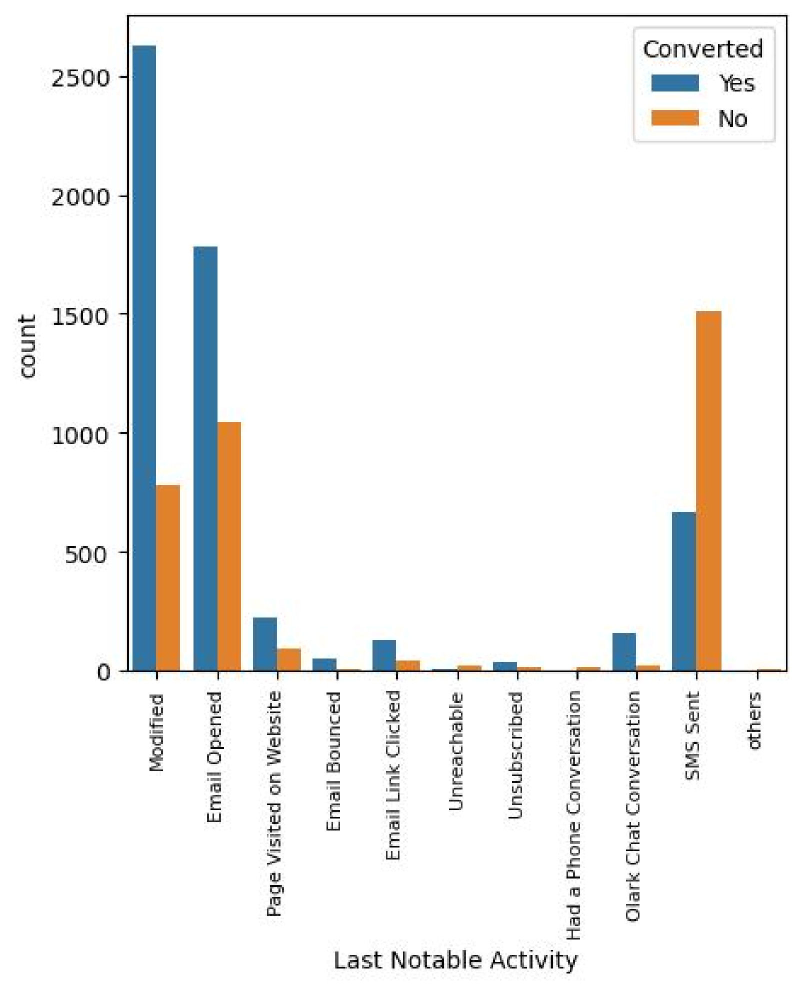




EDA

Categorical Variable Analysis





EDA

Catagorical Variable Analysis

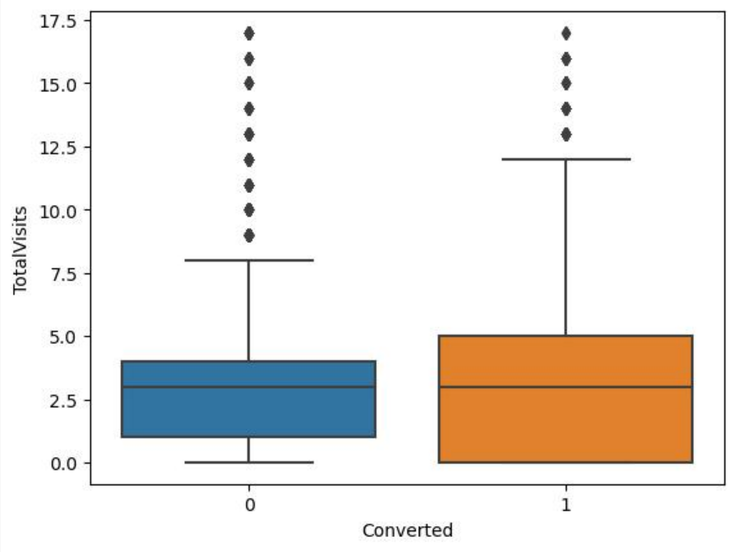
Observations:

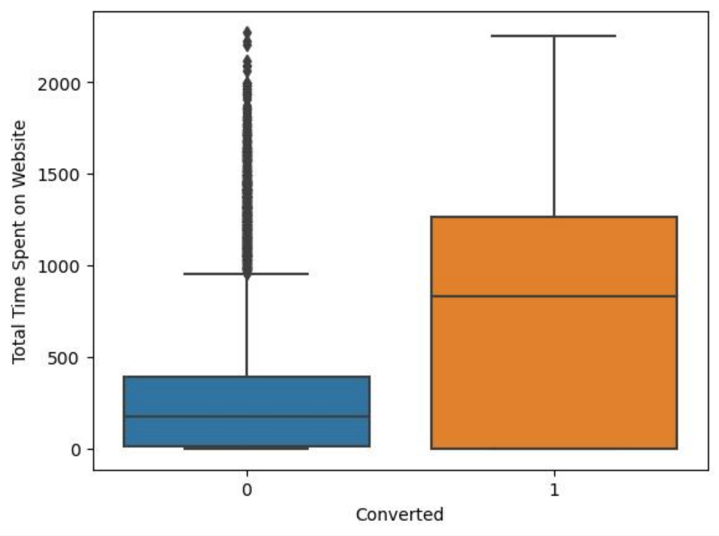
More conversions has been made when,

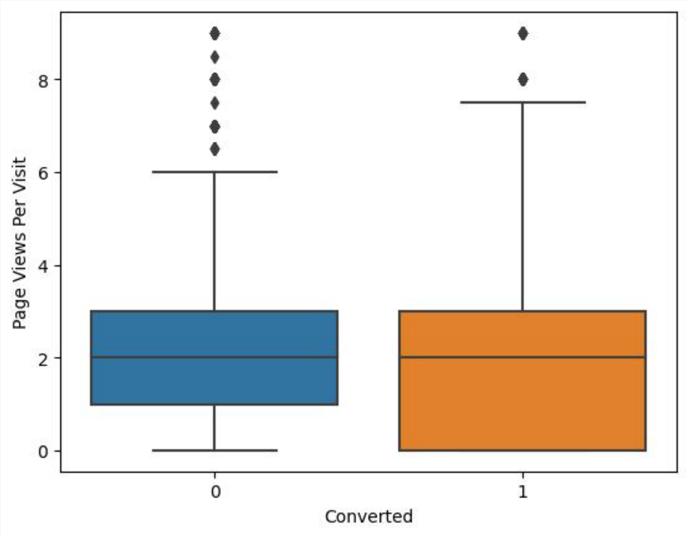
- Lead Origin is Landing Page Submission
- Lead Source is Google and Direct Traffic
- Last Activity is Email Opened
- What is your current occupation is Unemployed and Working Professionals
- Last Notable Activity is Modified

EDA

Numerical Variable Analysis







EDA

Numerical Variable Analysis

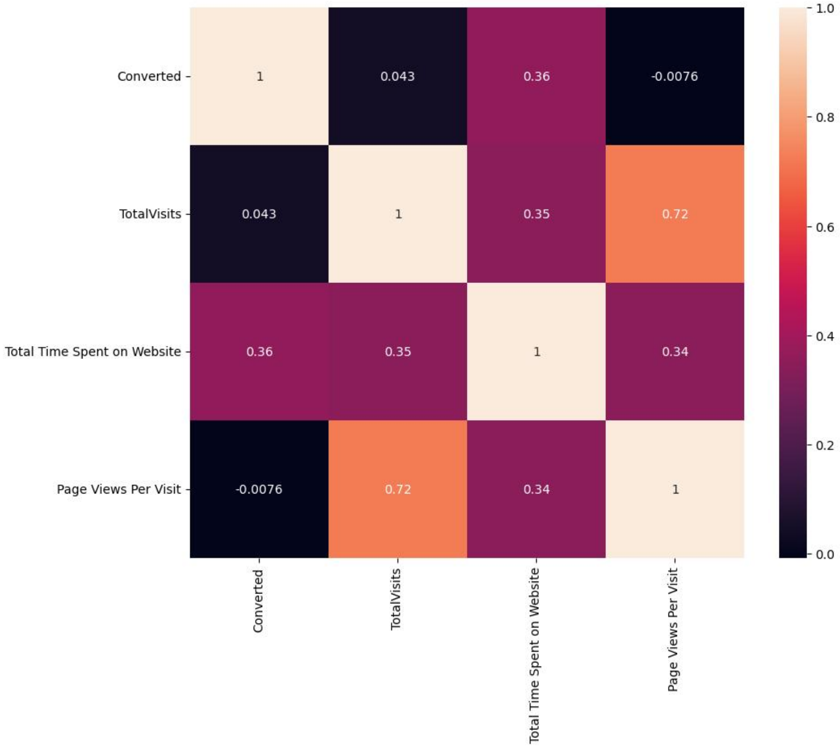
Observation:

- The median for both lead converted and lead non-converted are the same. So we cannot have any insights from this variable.
- the people who have spent more time in the website are likely to become the leads.
- The median for both lead converted and lead non-converted are the same. So we cannot have any insights from this variable.

CORRELATION MATRIX

The correlation matrix for the numerical variables with respect to the target variable:

- From the heatmap, we can see that
- The variables 'TotalVisits' and 'Total Time Spent on Website' are having positive correlation with the target variable.
- The variable 'Page Views Per Visit' is having a negative correlation with the target variable.
- The correlation coefficients are
- TotalVisits - 0.043
- Total Time Spent on Website' - 0.36
- Page Views Per Visit - -0.0076.



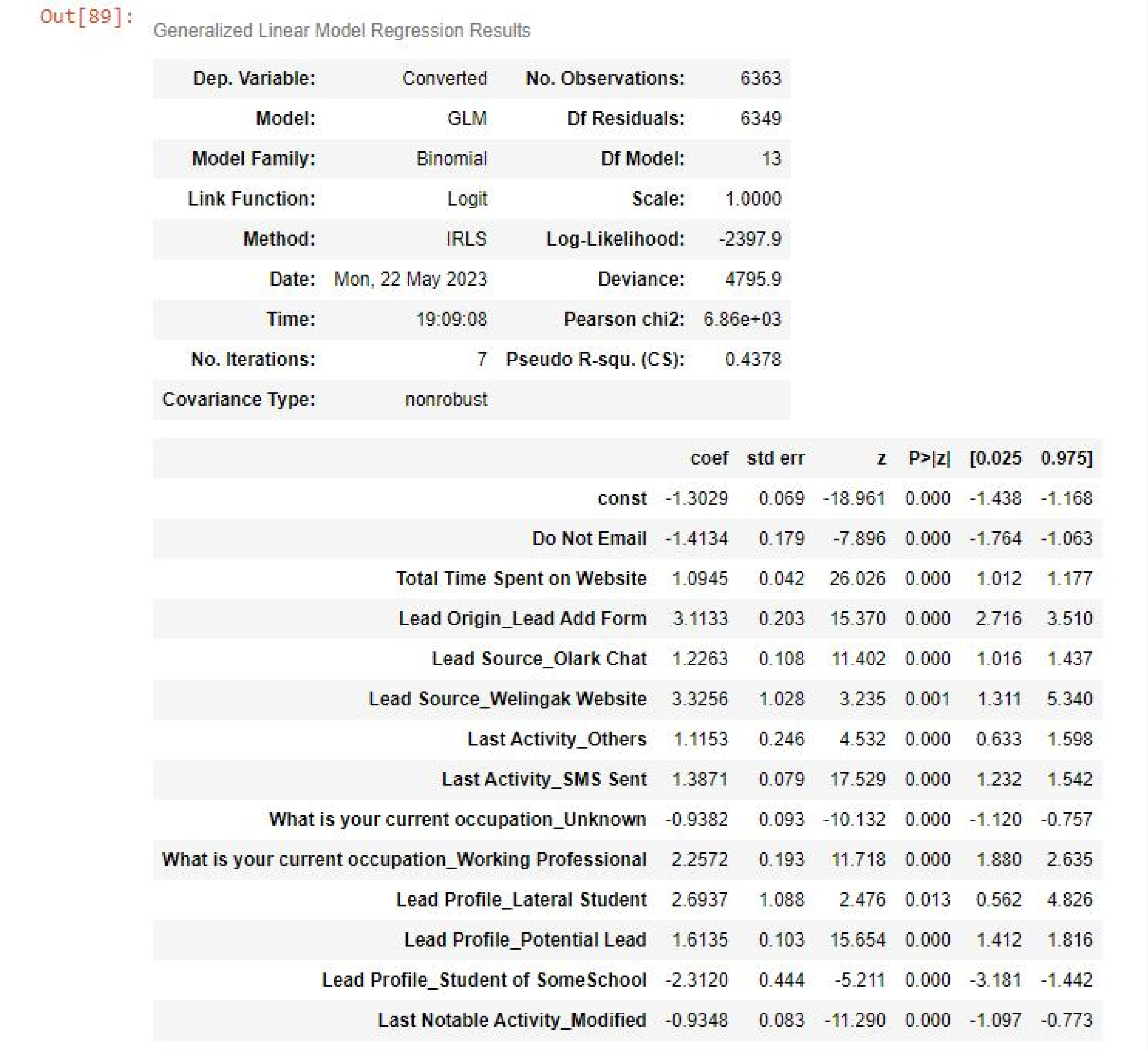
MODEL BUILDING

- **Data Preparation**
  - The binary categorical variables with Yes/No is converted to binary numerical variables with 1/0.
  - Dummy variables has been created for the categorical variables with multiple levels.
- **Train-Test Split**
  - The entire dataset has been split into training and testing data in the ratio 70:30 respectively.
- F**eature Scaling**
  - The variables in the dataset are changed to comparable scale using the Standardization method.
- **Model Building**

– The top 15 features are selected using the RFE method and model is build by removing the unwanted variables manually by checking the p-values and VIFs. A model with 13 variables whose p-values and VIFs are within range is created. Using this model, the probabilities for the lead to be converted as hot leads is calculated. A predicted value of 0 or 1 is obtained by using the threshold value as 0.5. A confusion matrix is obtained by using the predicted values. Then the model is evaluated by the metrics Accuracy, Sensitivity, Specificity, Precision and Recall. Then an optimal threshold value was obtained by using the Accuracy, Sensitivity and Specificity graph which is 0.3. On using the threshold 0.3, we obtained a model with a lead conversion rate of about 84%.

MODEL BUILDING

Model summary of the final model:



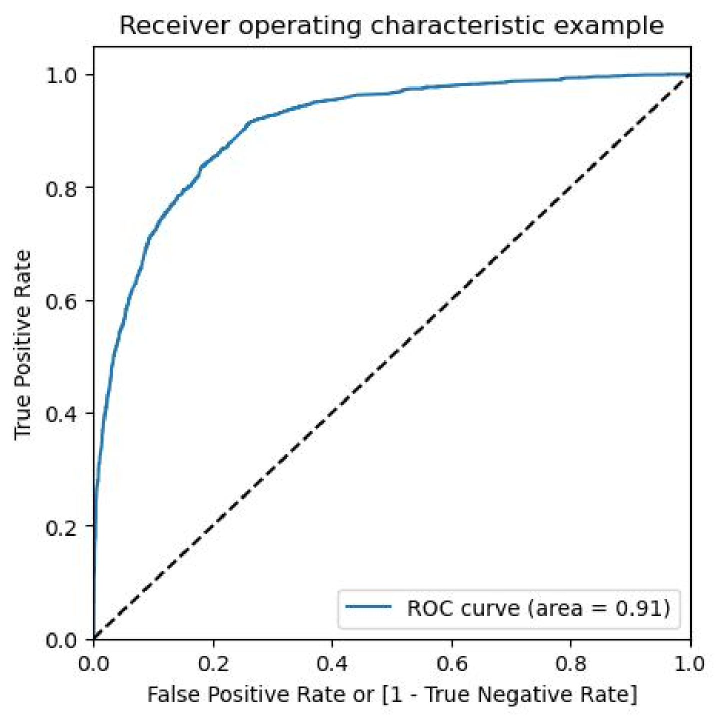
```
Out[89]:
         Generalized Linear Model Regression Results

     Dep. Variable:        Converted   No. Observations:      6363
            Model:              GLM       Df Residuals:      6349
     Model Family:         Binomial           Df Model:        13
    Link Function:            Logit              Scale:    1.0000
           Method:             IRLS     Log-Likelihood:   -2397.9
             Date: Mon, 22 May 2023           Deviance:    4795.9
             Time:         19:09:08      Pearson chi2:   6.86e+03
   No. Iterations:                7  Pseudo R-squ. (CS):   0.4378
  Covariance Type:        nonrobust

                                              coef  std err        z  P>|z|  [0.025  0.975]
                                   const   -1.3029    0.069  -18.961  0.000  -1.438  -1.168
                            Do Not Email   -1.4134    0.179   -7.896  0.000  -1.764  -1.063
               Total Time Spent on Website  1.0945    0.042   26.026  0.000   1.012   1.177
                Lead Origin_Lead Add Form   3.1133    0.203   15.370  0.000   2.716   3.510
                   Lead Source_Olark Chat   1.2263    0.108   11.402  0.000   1.016   1.437
             Lead Source_Welingak Website   3.3256    1.028    3.235  0.001   1.311   5.340
                    Last Activity_Others    1.1153    0.246    4.532  0.000   0.633   1.598
                  Last Activity_SMS Sent    1.3871    0.079   17.529  0.000   1.232   1.542
      What is your current occupation_Unknown  -0.9382  0.093  -10.132  0.000  -1.120  -0.757
What is your current occupation_Working Professional  2.2572  0.193  11.718  0.000  1.880  2.635
             Lead Profile_Lateral Student   2.6937    1.088    2.476  0.013   0.562   4.826
              Lead Profile_Potential Lead   1.6135    0.103   15.654  0.000   1.412   1.816
      Lead Profile_Student of SomeSchool   -2.3120    0.444   -5.211  0.000  -3.181  -1.442
           Last Notable Activity_Modified  -0.9348    0.083  -11.290  0.000  -1.097  -0.773
```

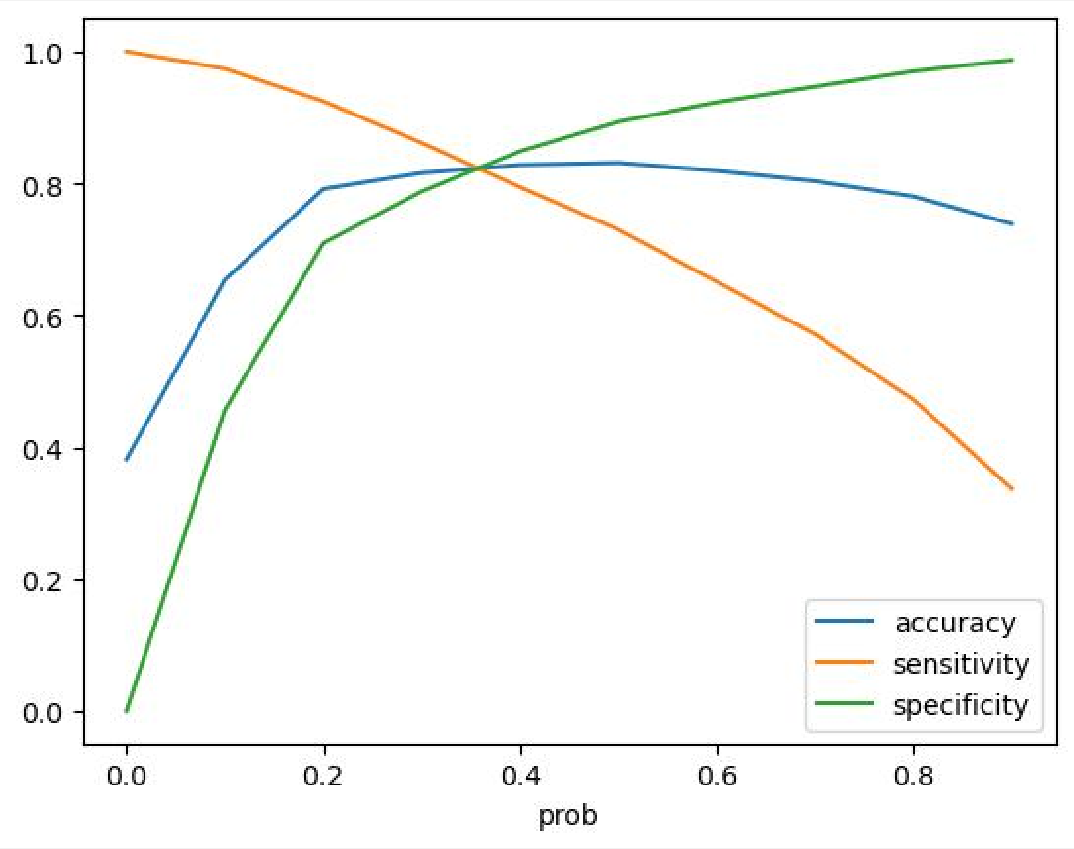
PLOTTING ROC

- We have obtianed the ROC curve with the area 0.91 which indicates that our model is more accurate.



FINDING OPTIMAL THRESHOLD

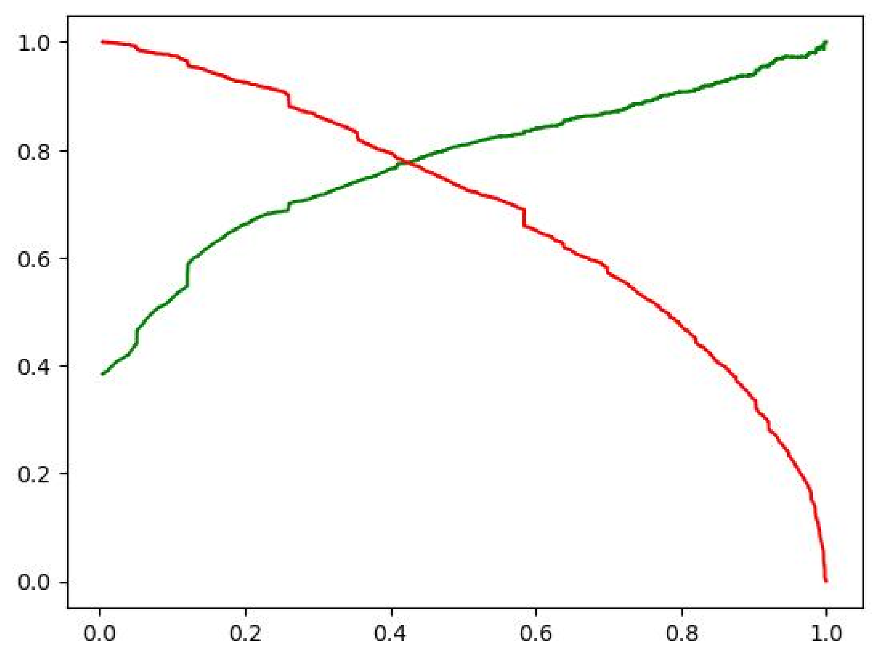
Accuracy, Sensitivity and Specificity graph



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

PRECISION-RECALL TRADEOFF

- The graph below represents the Precision-Recall tradeoff.



- Based on the Business problem we can either choose the "Sensitivity and Specificity tradeoff" or "precision and recall tradeoff"

- As our problem statement is to increase the lead converation rate we want our model to predict the 1's (Lead getting converted to paying customer) so we are choosing optimum cutoff value based on the sensitivity and specificity tradeoff.

MODEL EVALUATION

- **On applying the model on the training set, we have the following evaluation metrics**.
  - Accuracy - 0.82
  - Sensitivity - 0.86
  - Specificity - 0.79
  - Precision - 0.71
  - Recall - 0.86

- **On applying the model on the test set, we have the following evaluation metrics.**
  - Accuracy - 0.81
  - Sensitivity - 0.84
  - Specificity - 0.78
  - Precision - 0.71
  - Recall - 0.85

- **Lead Conversion Rate:**
  - The target lead conversion rate to be around 85% as expected by the CEO of X Education.

CONCLUSION AND RECOMMENDATION

- The features in our final model are listed with their coefficients as follows.
  - Do Not Email : -1.4133
  - Total Time Spent on Website : 1.0944
  - Lead Origin_Lead Add Form : 3.1132
  - Lead Source_Olark Chat : 1.2263
  - Lead Source_Welingak Website : 3.3256
  - Last Activity_Others : 1.1153
  - Last Activity_SMS Sent : 1.3870
  - What is your current occupation_Unknown : -0.9382
  - What is your current occupation_Working Professional : 2.2571
  - Lead Profile_Lateral Student : 2.6936
  - Lead Profile_Potential Lead : 1.6135
  - Lead Profile_Student of SomeSchool : -2.3119
  - Last Notable Activity_Modified : -0.9347
- We can increase the potential leads by concentrating on the above features.