LEAD SCORING CASE STUDY

ANUPRIYA KHARINTA

ANURAG AGARWAL

ANURAG RAJ

THE PROBLEM

- 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.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. 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%.

GOALS OF THE CASE STUDY

- 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

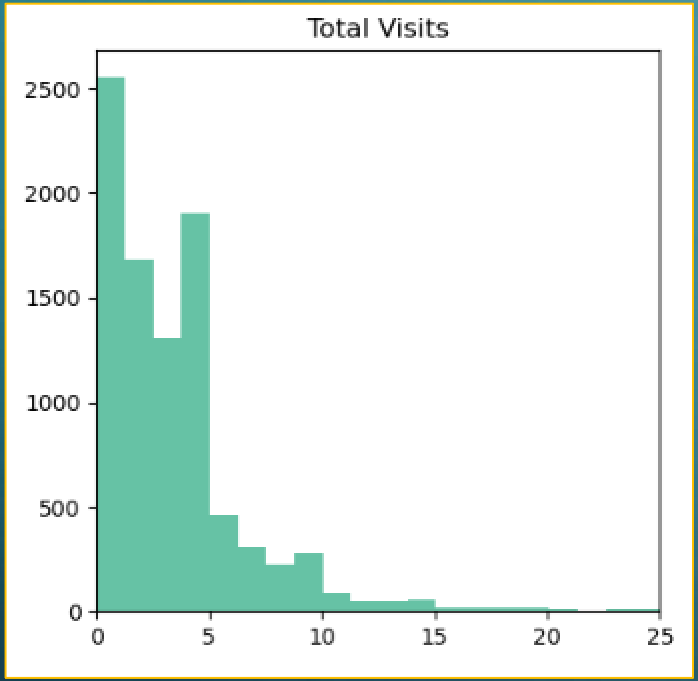
SOLUTIONS APPROACH

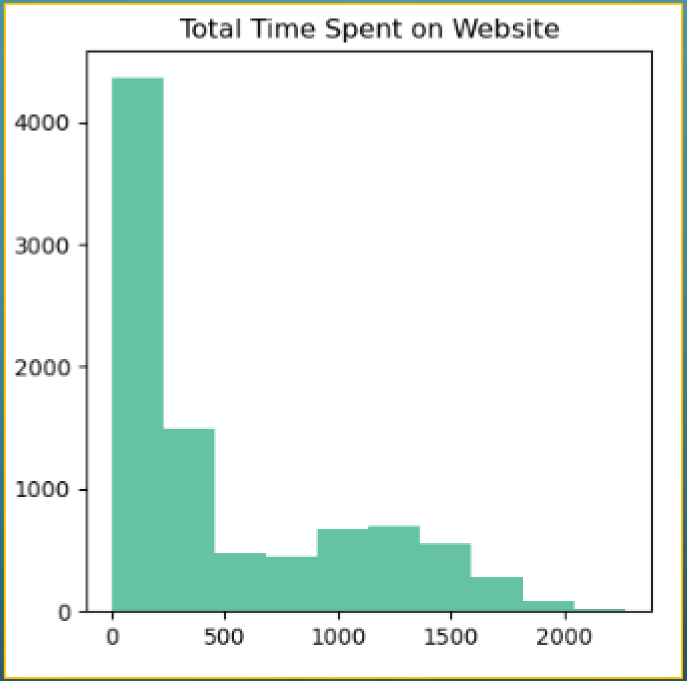
- Reading the Data
- Data Cleanup
- EDA and Visualizing Data
- Scaling features and creating dummy variables to encode the data.
- Splitting the data into training and test sets.
- Analysing correlations between variables.
- Building the model (including RFE, R-squared, VIF, and p-values).
- Evaluating the model's performance.
- Making predictions on the test set.

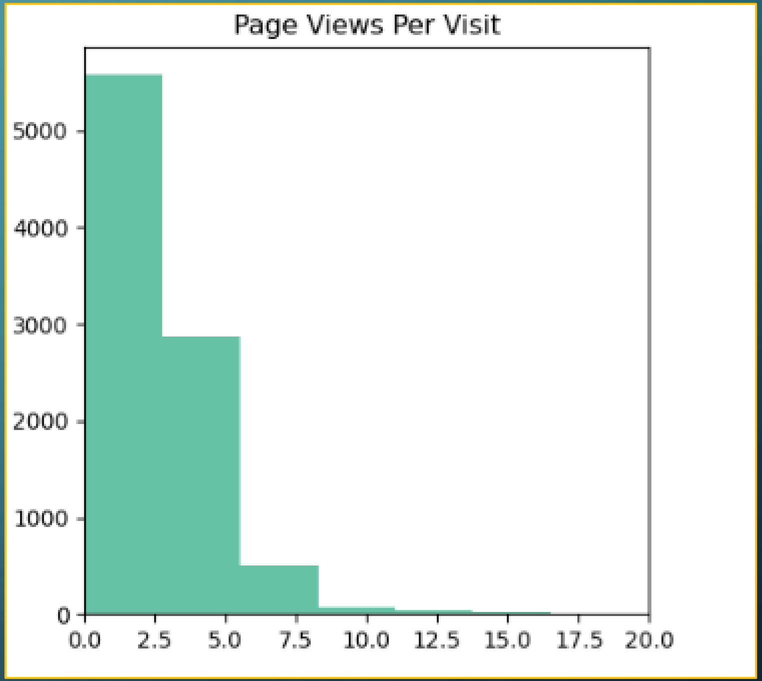
READING THE DATA & DATA CLEANUP

- Total Number of Rows =**37**, Total Number of Columns =**9240**.
- Since "Select" is not a valid category, it is likely the default value set in the form dropdown. If users didn't choose an option, the value remained as "Select." Therefore, we replaced "Select" with NaN.
- Dropped 'Magazine','Receive More Updates About Our Courses','I agree to pay the amount through cheque','Get updates on DM Content','Update me on Supply Chain Content’ as **all the values are same.**
- Dropped 'Asymmetrique Profile Index','Asymmetrique Activity Index','Asymmetrique Activity Score','Asymmetrique Profile Score','Lead Profile','Tags','Lead Quality','How did you hear about X Education','City','Lead Number’ as **more than 35% values are null.**
- Since the "Prospect ID" and "Lead Number" contain unique values and don't contribute to the analysis, they can be excluded from the dataset to streamline the process.

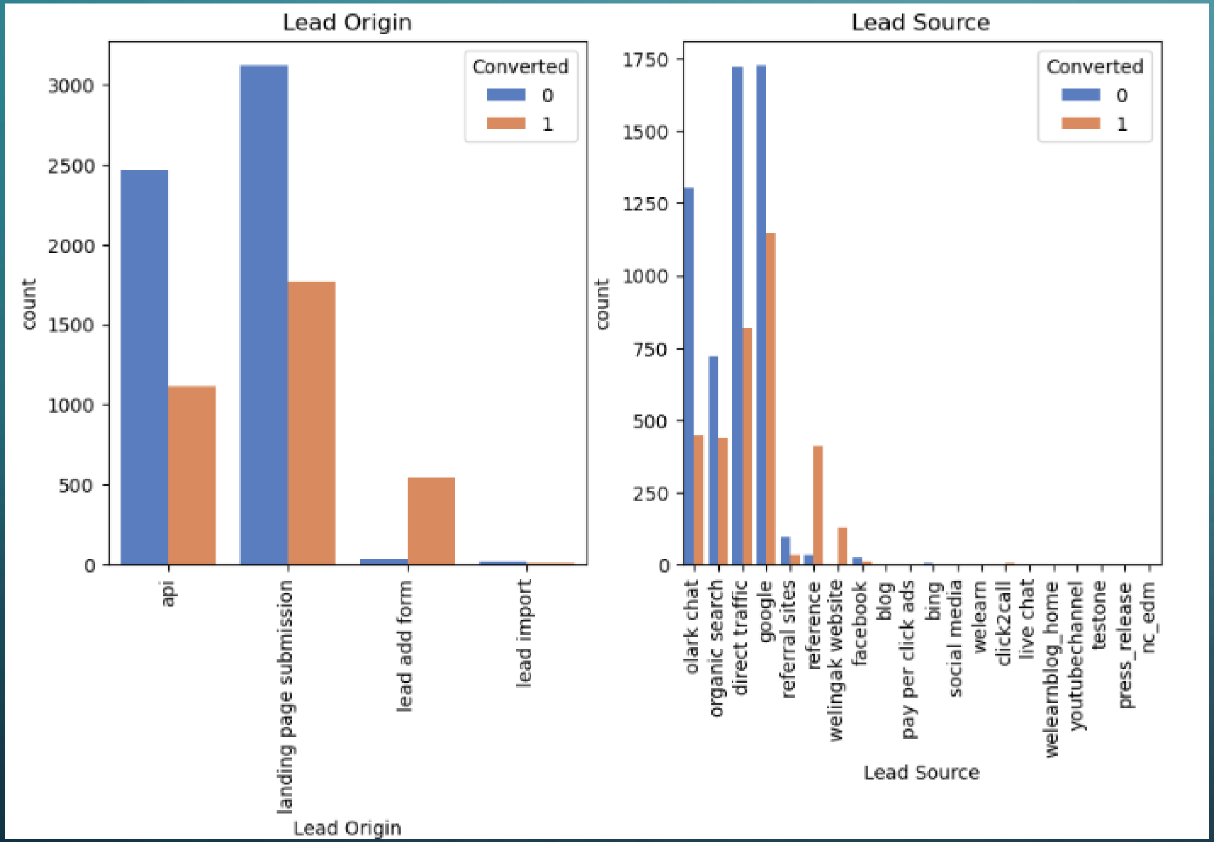
EDA AND VISUALIZING DATA : NUMERICAL VARIABLE


Total Visits


Total Time Spent on Website


Page Views Per Visit

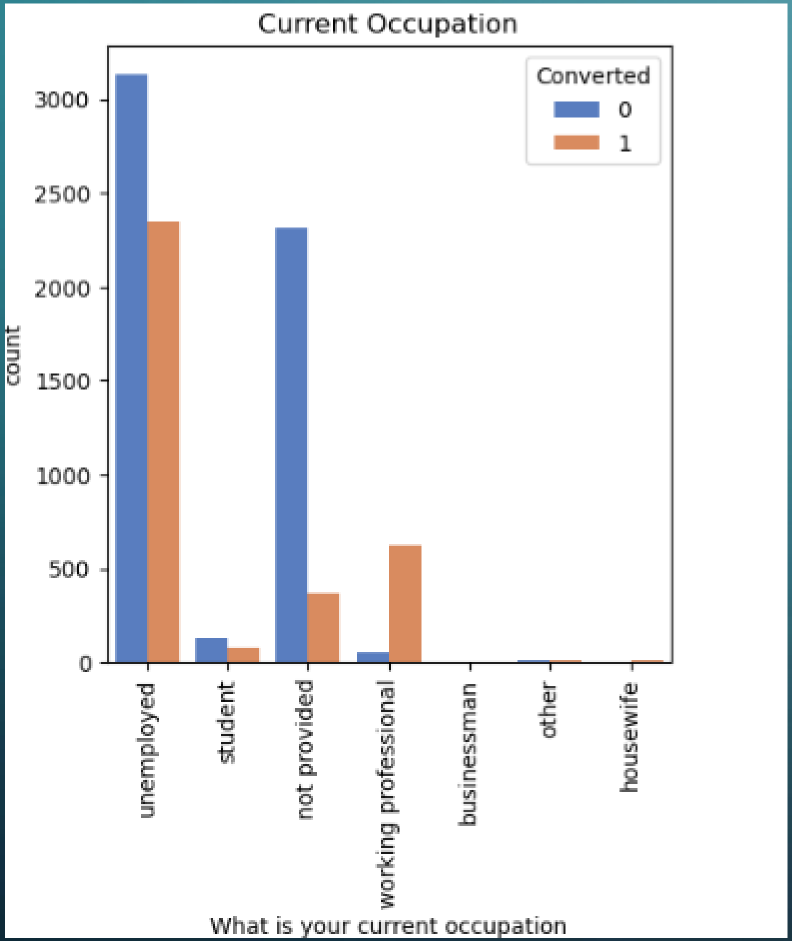
EDA AND VISUALIZING DATA : CATEGORICAL VARIABLE



- Leads originating from **Google** and **direct traffic** show a high probability of successful conversion.

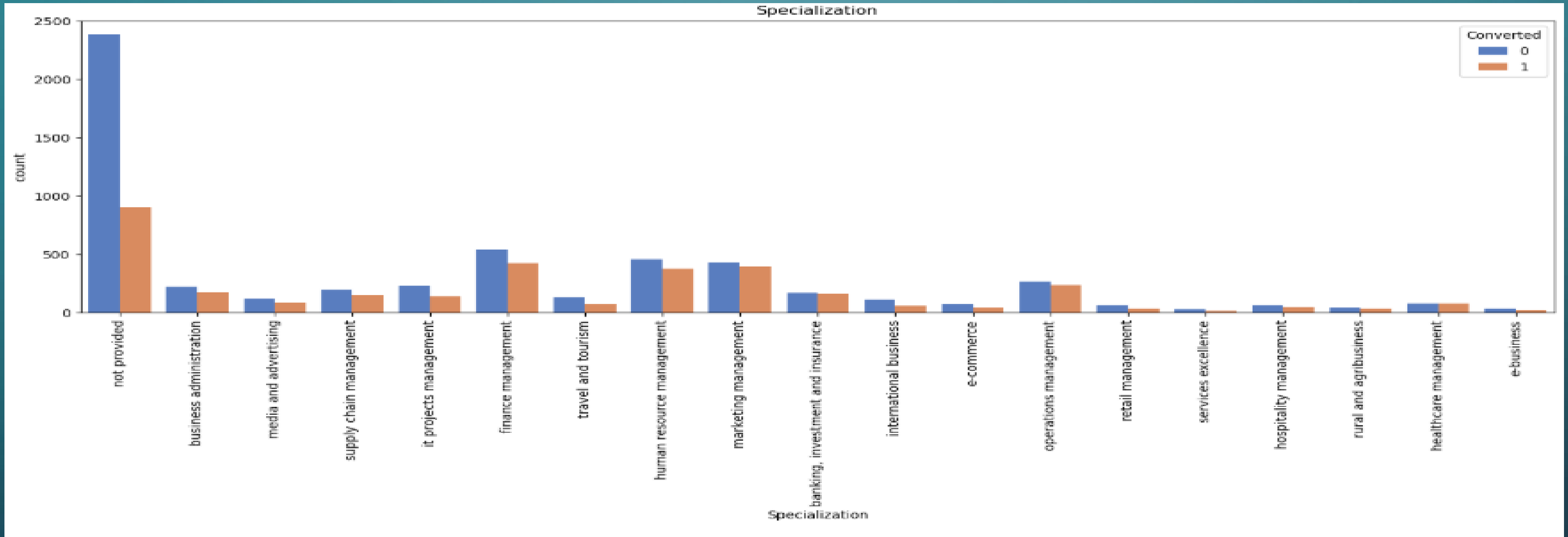
- Leads originating from "**Lead add form**" categories have a significantly higher likelihood of successful conversion.

EDA AND VISUALIZING DATA : CATEGORICAL VARIABLE



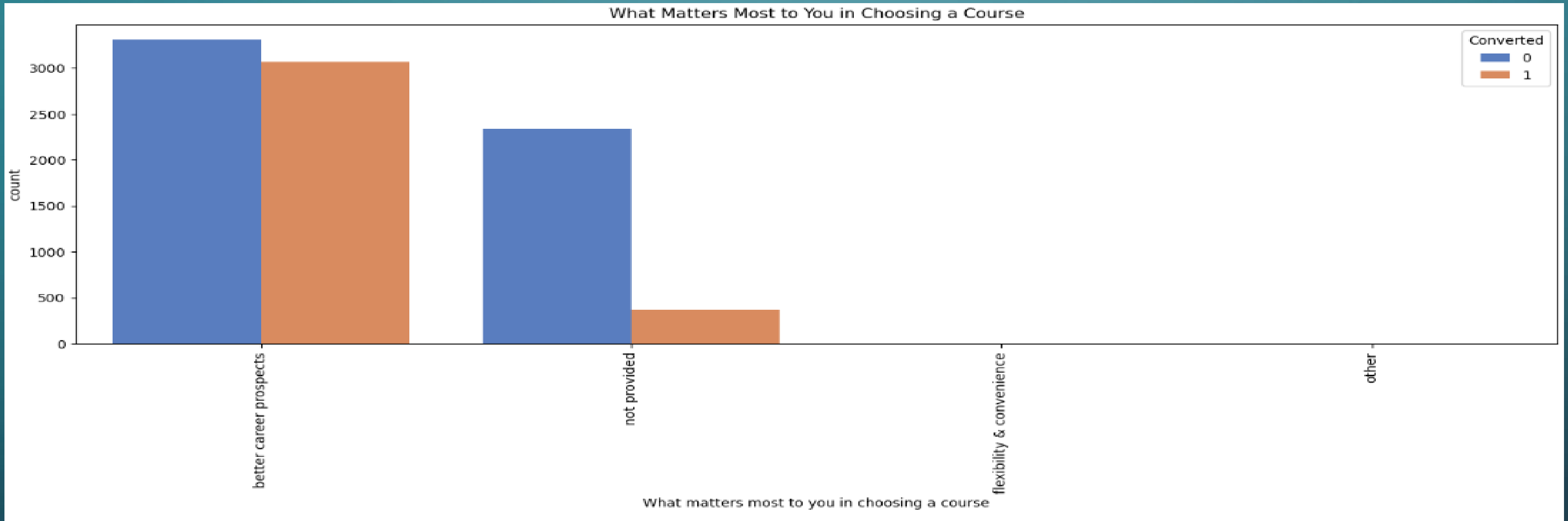
- Unemployed leads show a higher interest in joining the course compared to other leads.

EDA AND VISUALIZING DATA : CATEGORICAL VARIABLE



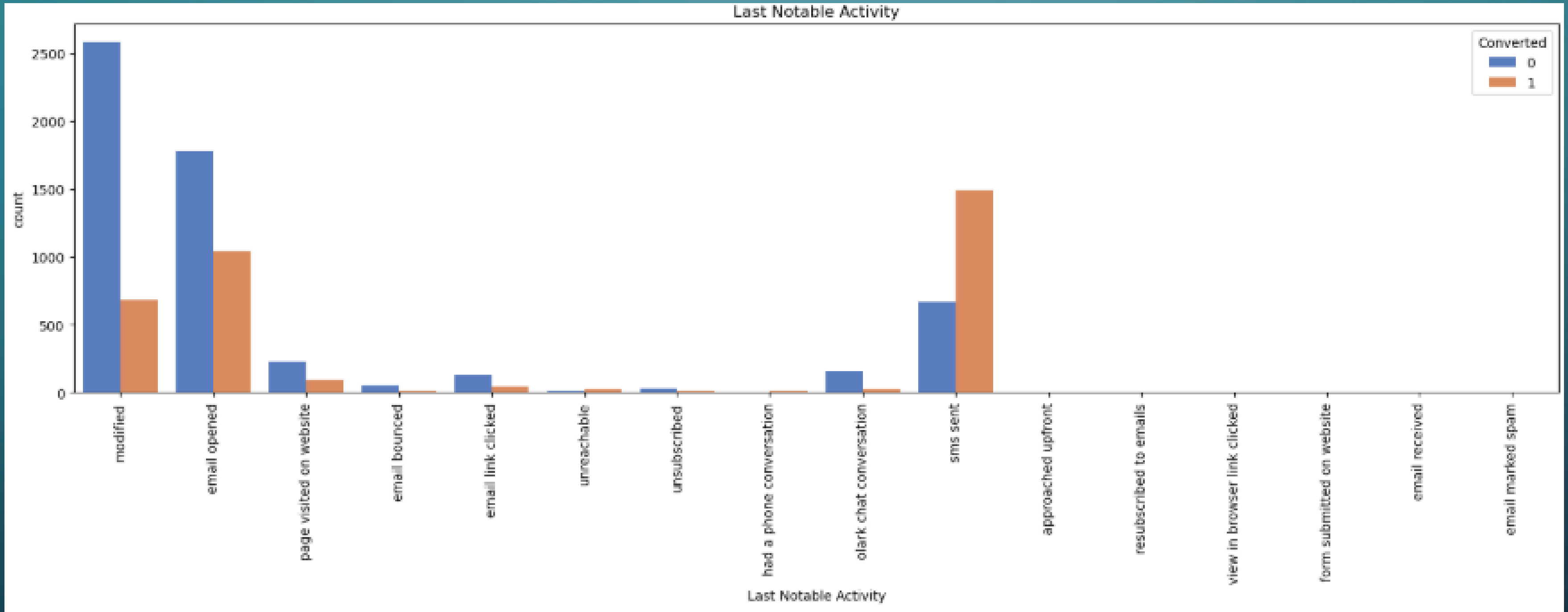
- Leads with specializations in HR, Finance, and Marketing Management exhibit a high probability of conversion.

EDA AND VISUALIZING DATA : CATEGORICAL VARIABLE



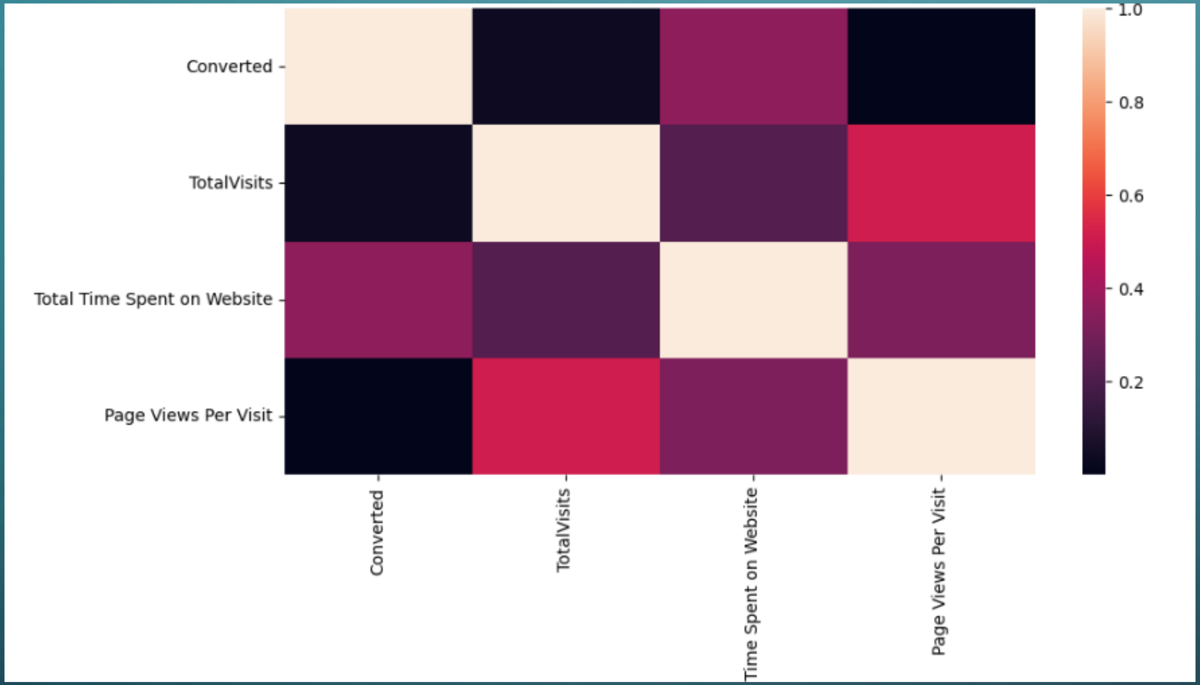
- Most of the leads are seeking courses to enhance their career prospects.

EDA AND VISUALIZING DATA : CATEGORICAL VARIABLE



- The highest conversion rate is associated with SMS sent.

CORRELATION



The variables do not exhibit any correlation with each other.

SCALING FEATURES AND CREATING DUMMY VARIABLES TO ENCODE THE DATA.

- Applied MinMax Scaling (fit and transform) on the training data for all numeric predictors
  - 'TotalVisits', 'Total Time Spent on Website', and 'Page Views Per Visit'.
- The following columns originally had Yes/No values, which were replaced with 1 for Yes and 0 for No:
  - Do Not Email, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, A free copy of Mastering The Interview
- Dummy variables were created for the following columns, with the original variable and the first dummy variable for each column being dropped from the data frame:
  - 'Lead Origin','Specialization' ,'Lead Source', 'Do Not Email', 'Last Activity', 'What is your current occupation','A free copy of Mastering The Interview', 'Last Notable Activity'

SPLITTING THE DATA INTO TRAINING AND TEST SETS.

- Dataset has been splitted into Train and Test in 70:30 ratio. Train dataset is used to train the model and Test dataset to evaluate the model.

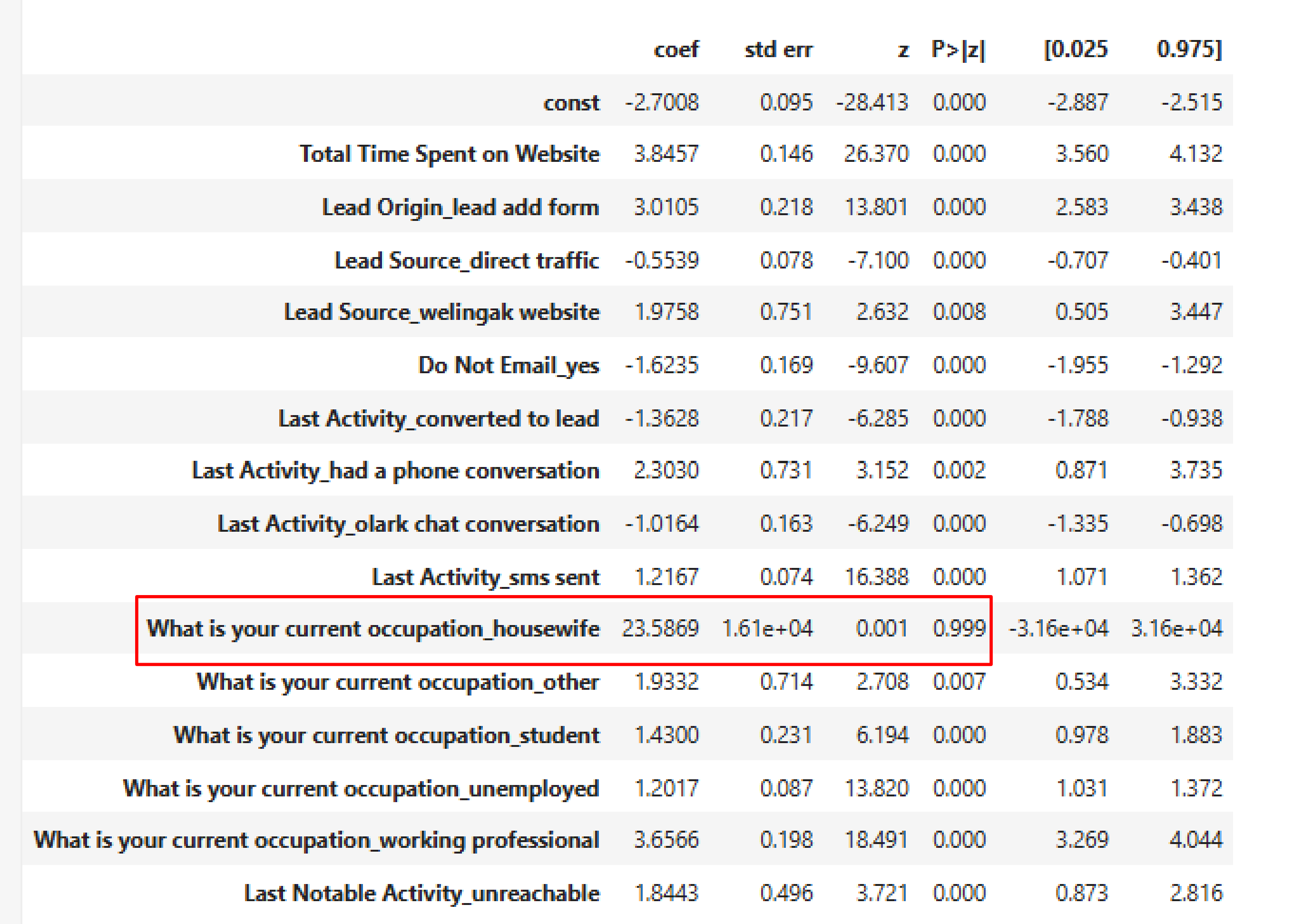
ANALYSING CORRELATIONS BETWEEN VARIABLES.

- **Lead Origin_lead import** has very high correlation with **Lead Source_facebook**.

BUILDING THE MODEL (INCLUDING RFE, R-SQUARED, VIF, AND P-VALUES).

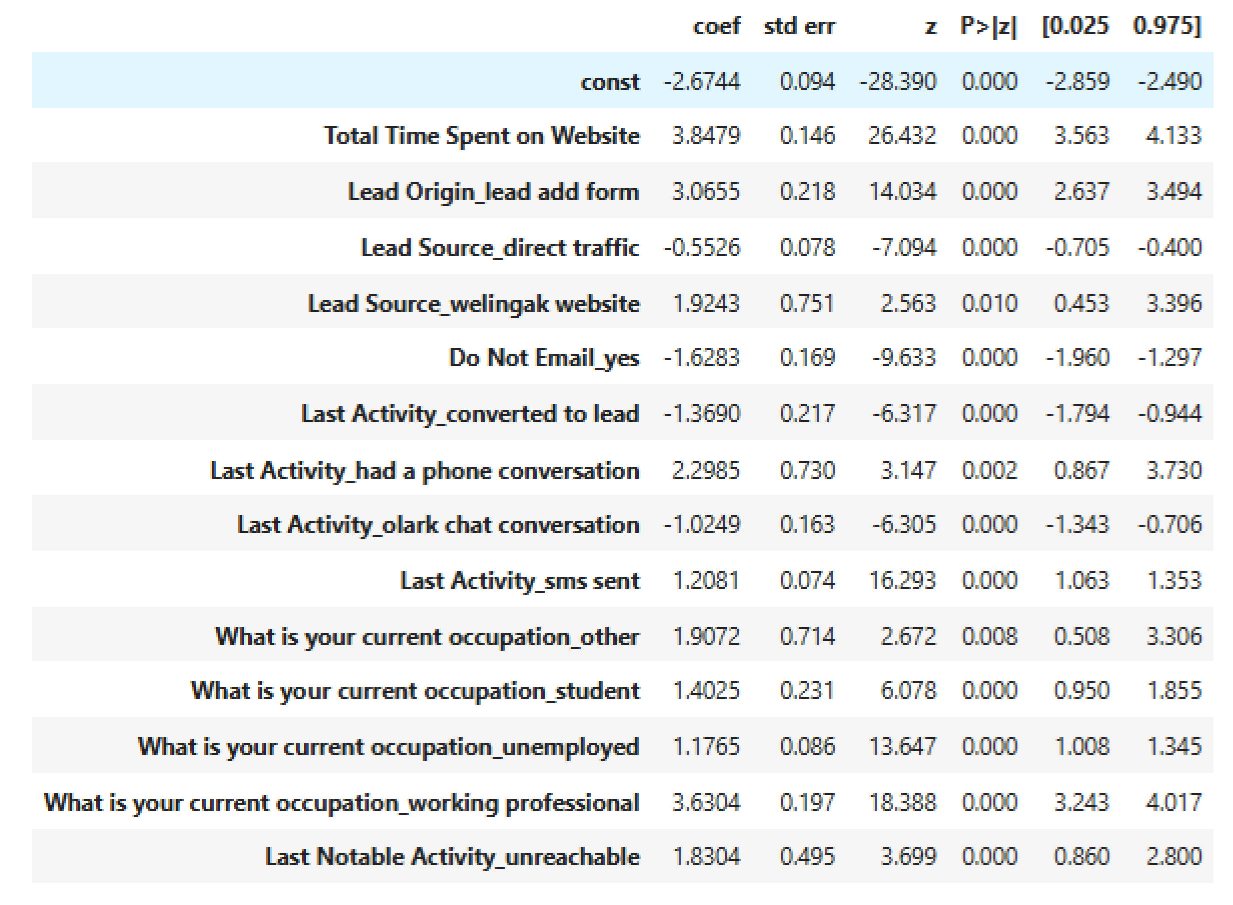
- Recursive Feature Elimination (RFE) has been used to get top 15 features.
- Building the model involves iteratively removing variables with a p-value greater than 0.05 and a VIF value greater than 5 to improve model accuracy and reduce multicollinearity.

BUILDING THE MODEL (INCLUDING RFE, R-SQUARED, VIF, AND P-VALUES).



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7008	0.095	-28.413	0.000	-2.887	-2.515
Total Time Spent on Website	3.8457	0.146	26.370	0.000	3.560	4.132
Lead Origin_lead add form	3.0105	0.218	13.801	0.000	2.583	3.438
Lead Source_direct traffic	-0.5539	0.078	-7.100	0.000	-0.707	-0.401
Lead Source_welingak website	1.9758	0.751	2.632	0.008	0.505	3.447
Do Not Email_yes	-1.6235	0.169	-9.607	0.000	-1.955	-1.292
Last Activity_converted to lead	-1.3628	0.217	-6.285	0.000	-1.788	-0.938
Last Activity_had a phone conversation	2.3030	0.731	3.152	0.002	0.871	3.735
Last Activity_olark chat conversation	-1.0164	0.163	-6.249	0.000	-1.335	-0.698
Last Activity_sms sent	1.2167	0.074	16.388	0.000	1.071	1.362
What is your current occupation_housewife	23.5869	1.61e+04	0.001	0.999	-3.16e+04	3.16e+04
What is your current occupation_other	1.9332	0.714	2.708	0.007	0.534	3.332
What is your current occupation_student	1.4300	0.231	6.194	0.000	0.978	1.883
What is your current occupation_unemployed	1.2017	0.087	13.820	0.000	1.031	1.372
What is your current occupation_working professional	3.6566	0.198	18.491	0.000	3.269	4.044
Last Notable Activity_unreachable	1.8443	0.496	3.721	0.000	0.873	2.816

BUILDING THE MODEL (INCLUDING RFE, R-SQUARED, VIF, AND P-VALUES).



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6744	0.094	-28.390	0.000	-2.859	-2.490
Total Time Spent on Website	3.8479	0.146	26.432	0.000	3.563	4.133
Lead Origin_lead add form	3.0655	0.218	14.034	0.000	2.637	3.494
Lead Source_direct traffic	-0.5526	0.078	-7.094	0.000	-0.705	-0.400
Lead Source_welingak website	1.9243	0.751	2.563	0.010	0.453	3.396
Do Not Email_yes	-1.6283	0.169	-9.633	0.000	-1.960	-1.297
Last Activity_converted to lead	-1.3690	0.217	-6.317	0.000	-1.794	-0.944
Last Activity_had a phone conversation	2.2985	0.730	3.147	0.002	0.867	3.730
Last Activity_olark chat conversation	-1.0249	0.163	-6.305	0.000	-1.343	-0.706
Last Activity_sms sent	1.2081	0.074	16.293	0.000	1.063	1.353
What is your current occupation_other	1.9072	0.714	2.672	0.008	0.508	3.306
What is your current occupation_student	1.4025	0.231	6.078	0.000	0.950	1.855
What is your current occupation_unemployed	1.1765	0.086	13.647	0.000	1.008	1.345
What is your current occupation_working professional	3.6304	0.197	18.388	0.000	3.243	4.017
Last Notable Activity_unreachable	1.8304	0.495	3.699	0.000	0.860	2.800

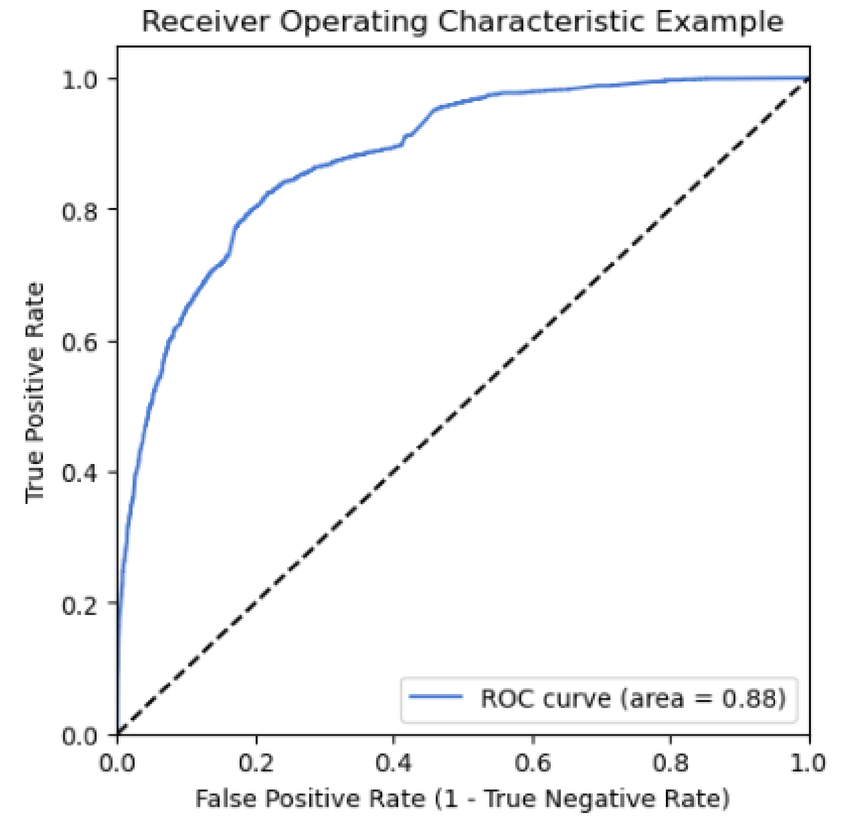
EVALUATING THE MODEL'S PERFORMANCE.

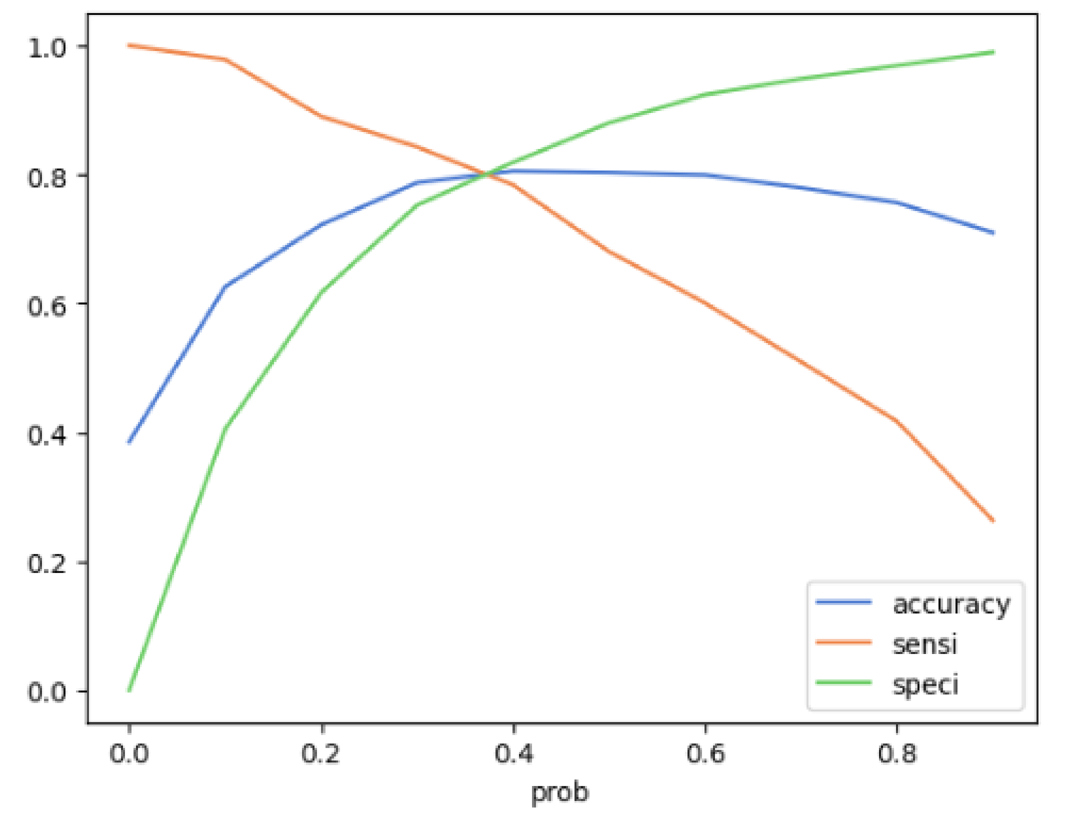
- After predicting the target on our training data set, we evaluated the model's performance using the following metrics:
  - Accuracy: 0.8030231459612659
  - Sensitivity: 0.6802943581357318
  - Specificity: 0.8798975672215109
  - Confusion Matrix:

    [3436, 469],

    [782, 1664]
  - ROC-AUC Score: The area under the ROC curve (AUC) is 0.88, indicating excellent performance.
- From the below graph In above plot, it’s visible that 0.35 is the optimal point to set as cutoff probability for our model.

EVALUATING THE MODEL'S PERFORMANCE.





MAKING PREDICTIONS ON THE TEST SET.

- Evaluating the model's performance on testing data set using the following metrics:
  - Accuracy: 0.8049944913698127
  - Sensitivity: 0.7997977755308392
  - Specificity: 0.8079584775086506
  - Confusion Matrix:

    ```
    [1401,  333],
    [ 198,  791]
    ```

With a current cutoff of 0.35, our accuracy, sensitivity, and specificity are approximately 80%.

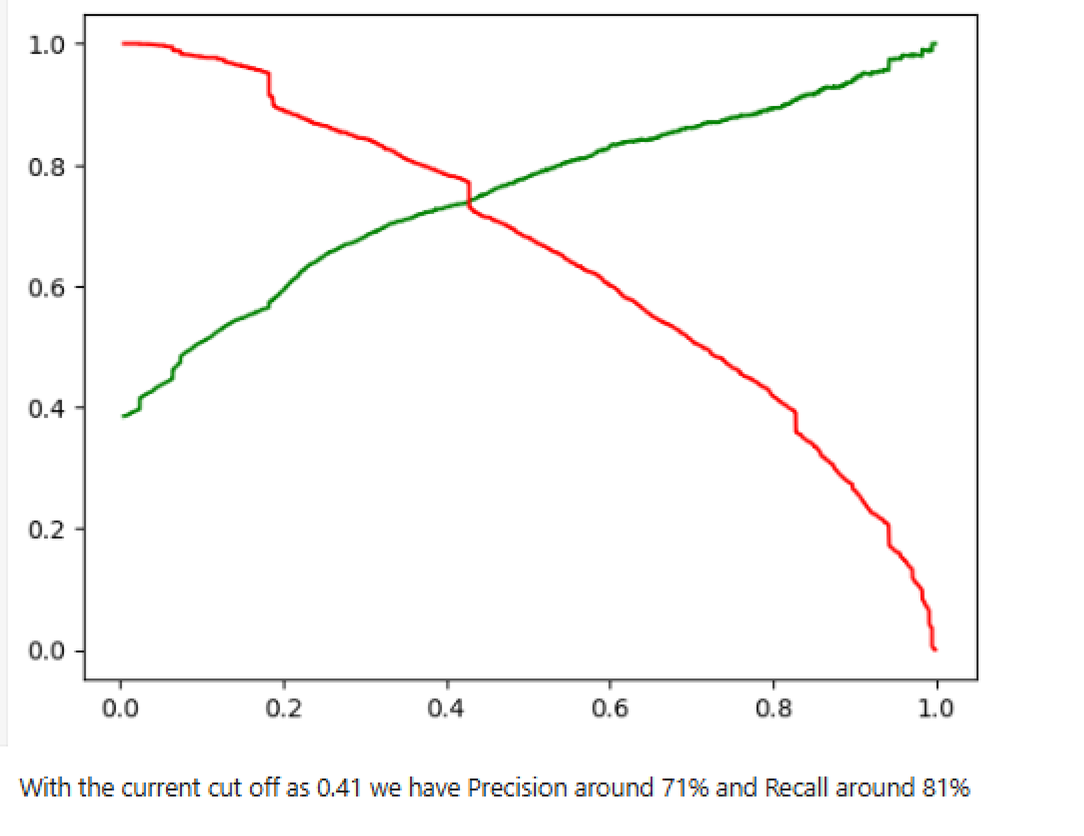
TRAINING MODEL UTILIZING THE PRECISION-RECALL PERSPECTIVE.

- After predicting the target on our training data set, we evaluated the model's performance using the following metrics:
  - Accuracy: 0.7995591245473154
  - Precision: 0.710290426676228
  - Recall: 0.8098937040065413
  - Confusion Matrix:

    [3097, 808],

    [465, 1987]
- From the below graph In above plot, it’s visible that the current cut off is 0.41 with Precision around 71% and Recall around 81%

TRAINING MODEL UTILIZING THE PRECISION-RECALL PERSPECTIVE.



With the current cut off as 0.41 we have Precision around 71% and Recall around 81%

MAKING FINAL PREDICTIONS ON THE TEST SET UTILIZING THE PRECISION-RECALL PERSPECTIVE

- Evaluating the model's performance on testing data set using the following metrics:
  - Accuracy: 0.8123393316195373
  - Precision: 0.7267552182163188
  - Recall: 0.7745197168857432
  - Confusion Matrix:

    ```
    [1446,  288],
    [ 223,  766]
    ```

With the current cut off as 0.41 we have Precision around 73% and Recall around 77%

CONCLUSION

- Prioritize contacting leads with higher Lead Scores first. Assign dedicated support to small batches of these hot leads to increase conversion chances.
- Unemployed leads show a higher interest in joining the course compared to other leads.
- The majority of leads are generated through Google and direct traffic channels. This indicates that these sources are highly effective in attracting potential customers to our website.
- Leads who spend more time on the website are more likely to convert into paying customers.
- The highest conversion rate is associated with SMS sent.