Lead Score Case Study

By: Kirti Garg

Problem Statement

- X-Education is an education company sells online Education courses to professionals and marketing through online advertisements. Company gets information through different channels and if candidates enquring with certain education level it calls lead. Typically lead conversion is 30% of certain education. Company identifying Hot Leads on certain criteria also. Lead conversion ratio is lesser than number of enrollment. comapny given Target to achieve 80% of total enrollment

Goal

- Building logistics regression model to finding leads for Company and help to achieve potential targets.
- Alternative approach should be ready in case Company's requirement changes in futures should be flexible.
-

Strategy

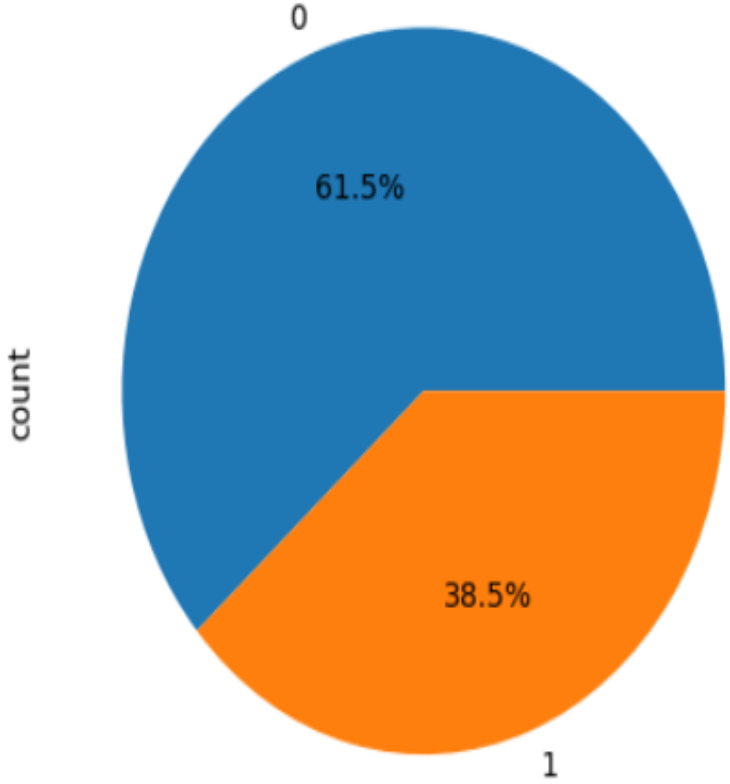
- Importing the necessary Libraries
- Data Loading
- Data Understanding
- Problem Statement Understanding
- Data Processing (Missing Value Imputation)
- EDA (Exploratory Data Analysis)
- Data Preprocessing (Dummy Variables, dropping unnecessry column,drop duplicates)
- Define X as predictors and y as target column
- Train-Test Split
- Scaling
- Feature Selection
- Modelling
- p-value check (p-values<0.05)
- VIF check (VIF <=5)
- Evaluation

Data Processing

- Calculated null values of all the columns.
- which has more than 30% of null values, removed those columns
- For the remaining numerical columns with Nan, imputed them with median
- For categorical columns with Nan, imputed them with mode.
- Handled Outliers as well

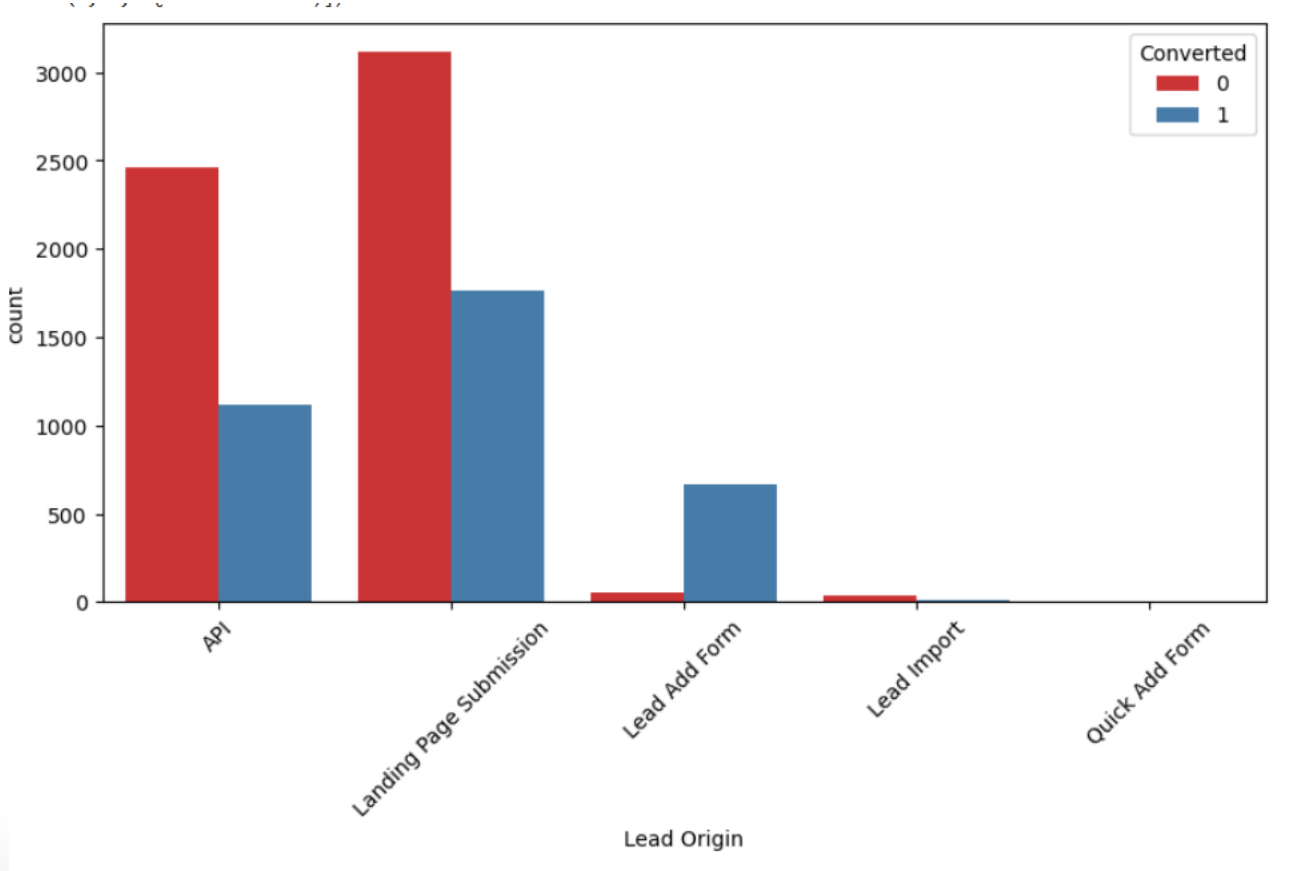
EDA (Exploratory Data Analysis)

According to the previous data Only 38.5% people have converted



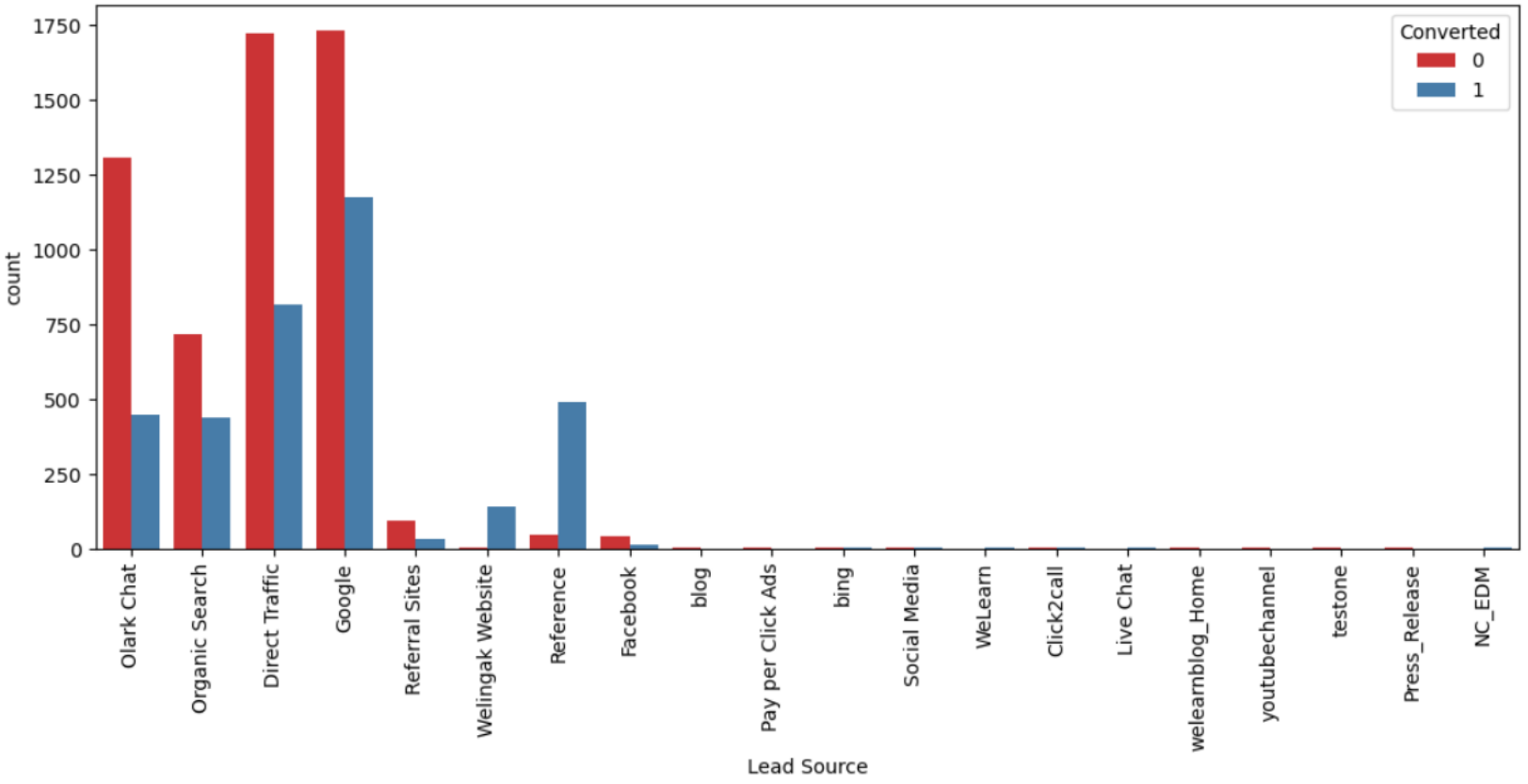
Lead Origin vs Converted

Land page submission has the higher conversion rate.



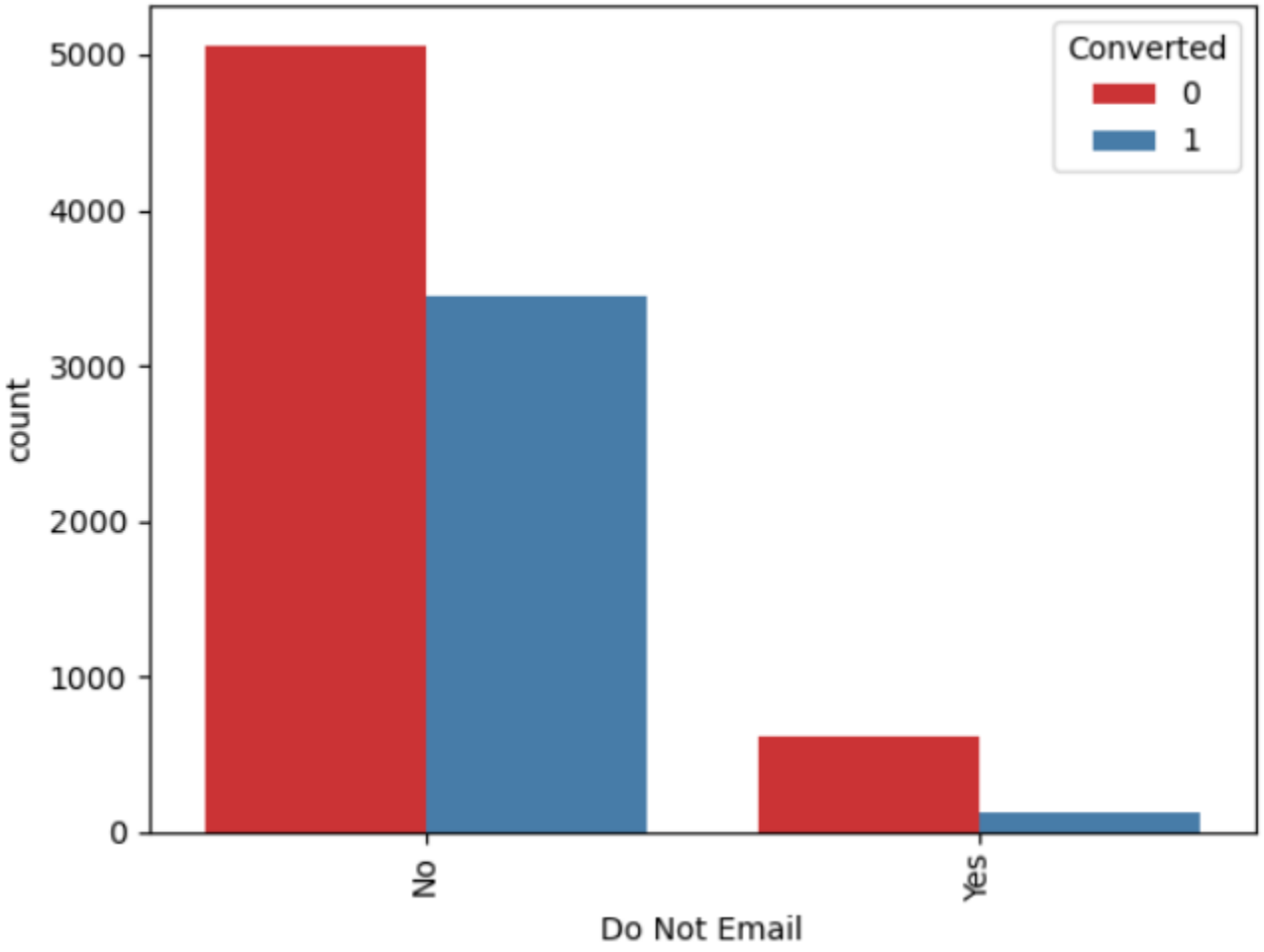
Lead Source vs Converted

people who opened the website from google have been converted the most



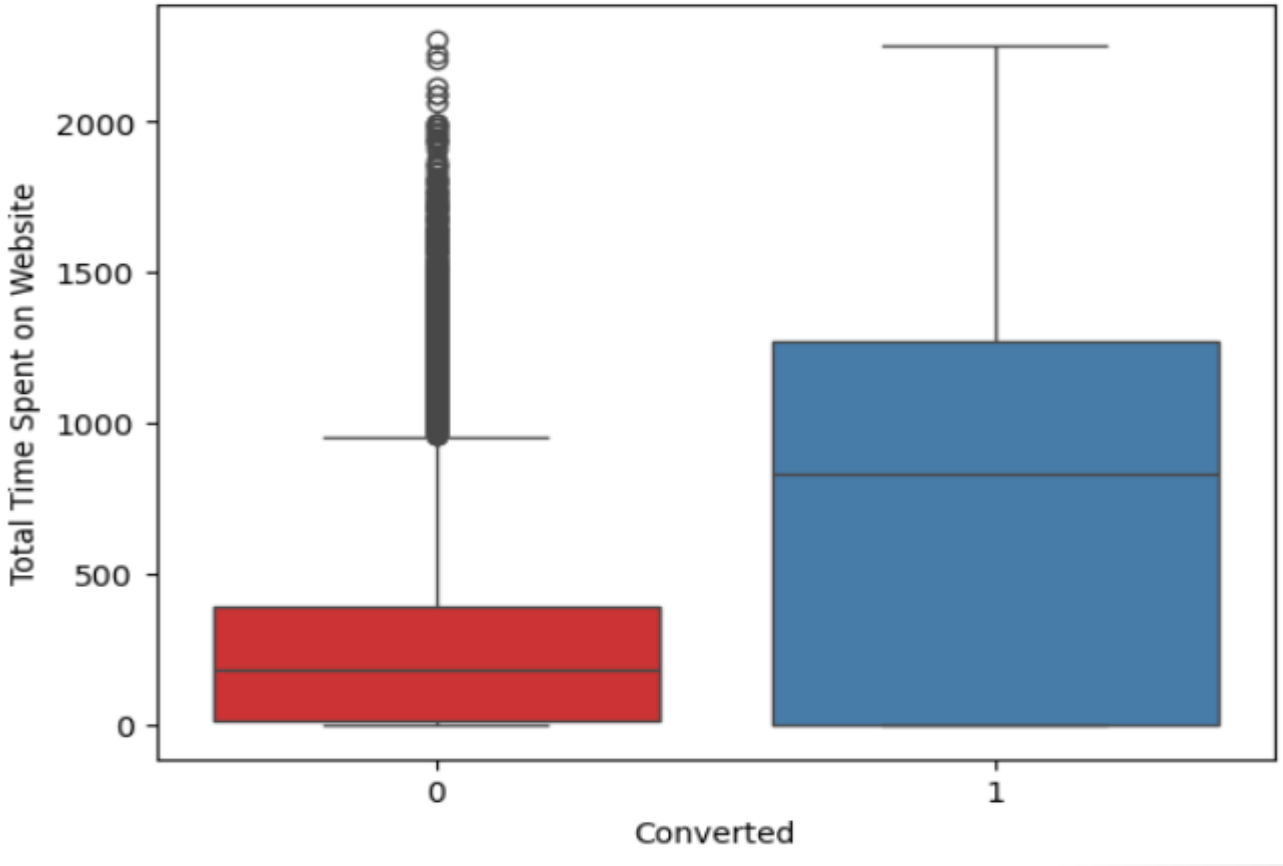
Do Not Email vs Converted

people who choose 'no' for do not email has more conversions



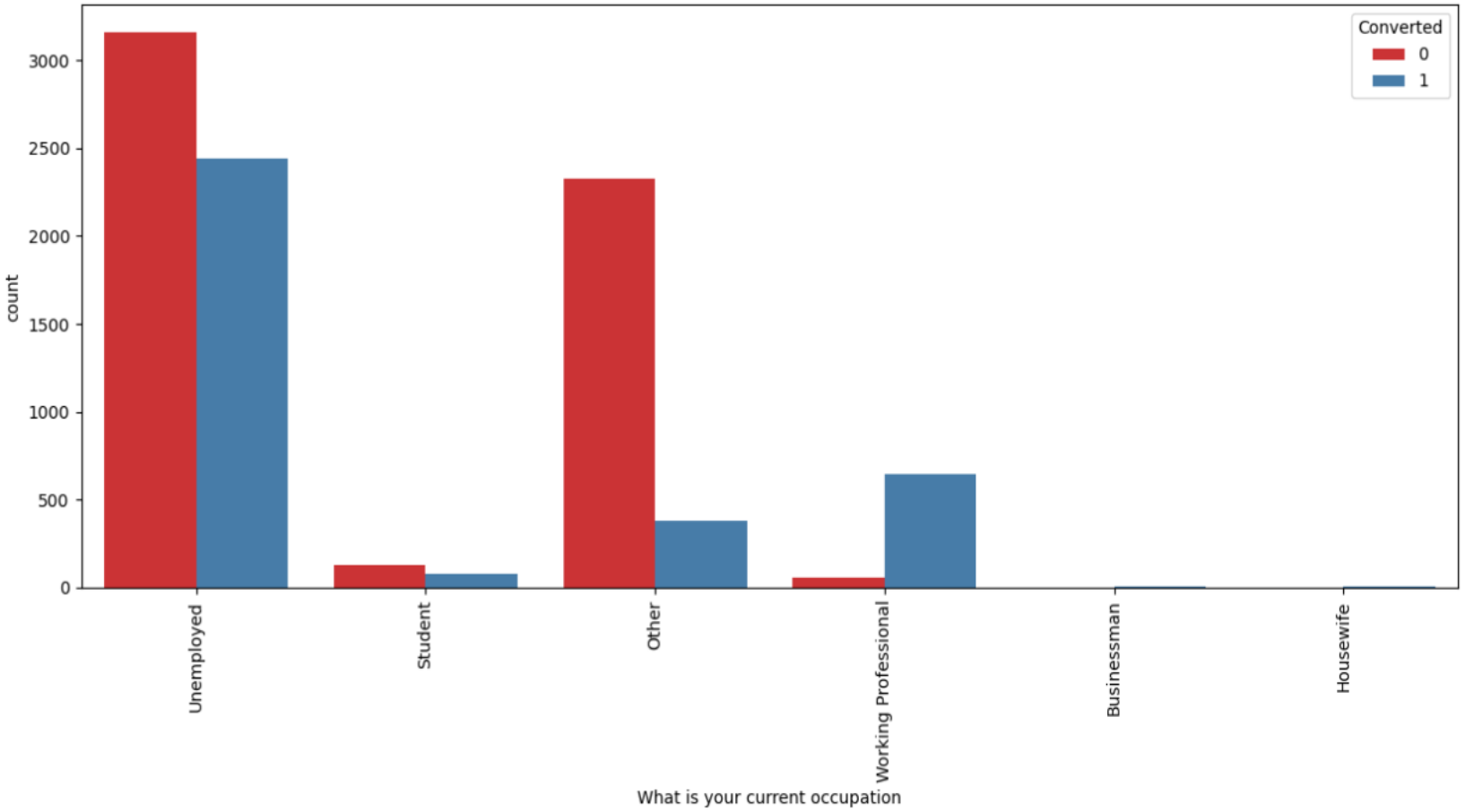
Total Time Spent on Website vs Converted

people who spent a lot of time on website has higher conversions



What is your current occupation vs Converted

unemployed people had converted the most



Data Preprocessing

- Firstly we saw which categorical column has yes or no values then changed them into binary values (0,1).
- For categorical variables with multiple levels, created dummy features (one-hot encoded).
- Dropped those columns whose dummies have been created.

Define X as predictors and y as target column

- Defined X variables which are all the column except converted column
- Y variables which is converted column.

Train-Test Split

- Splited Train and test data sets.
- So there will be 4 data sets, X_train, y_train, X_test and y_test
- Train sets for training the model and test set for checking the model and making predictions.

Scaling

- We do feature scaling so that it will not affect the model creation with different ranges.
- We took StandardScaler

First Model Creation

- Created first model with all the X variables to see the p values of all of them .
- To take the rough idea of the model

Feature Selection

- Used RFE for feature removal
- with RFE columns which are not ranked most affectable will be removed. impactable columns will be there
- Took 15 columns from RFE

Modelling

- Created the final model with p value less than 0.05 and VIF less than 5.
- There was some variables with p value more than 0.05 and then it was removed for the good model

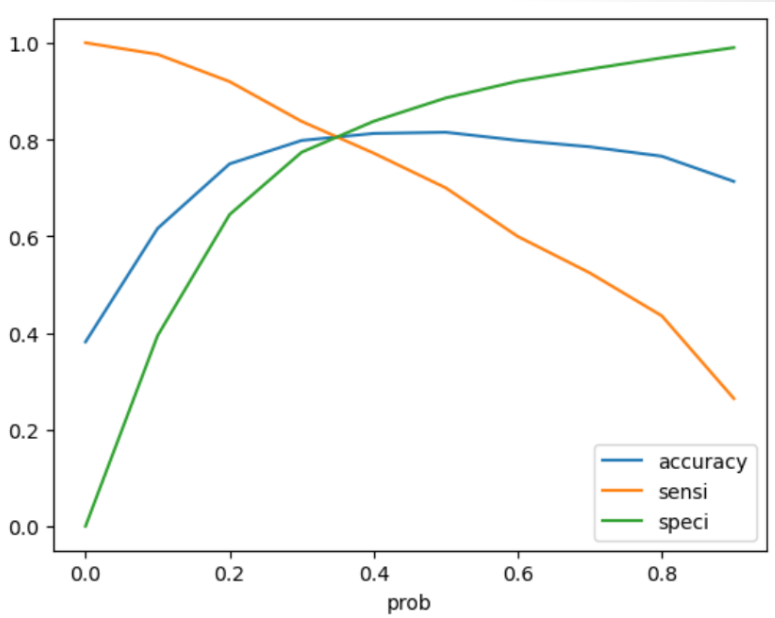
Generalized Linear Model Regression Results

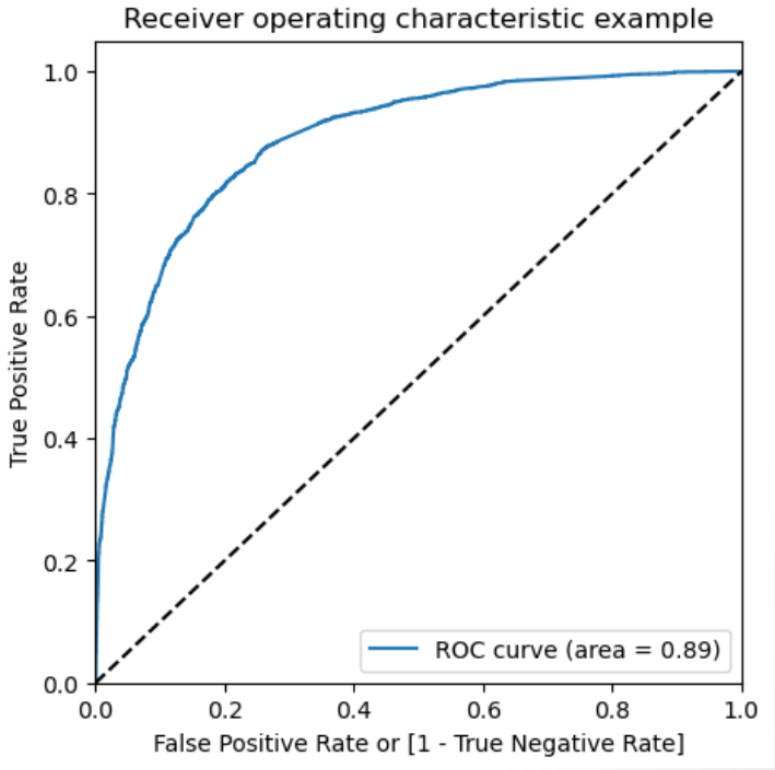
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6455
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2659.7
Date:	Sat, 14 Sep 2024	**Deviance:**	5319.4
Time:	02:39:25	**Pearson chi2:**	7.12e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3977
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3818	0.123	-3.106	0.002	-0.623	-0.141
Do Not Email	-1.2960	0.166	-7.794	0.000	-1.622	-0.970
Total Time Spent on Website	1.0715	0.040	27.020	0.000	0.994	1.149
Lead Origin_Landing Page Submission	-0.9609	0.127	-7.550	0.000	-1.210	-0.711
Lead Origin_Lead Add Form	2.9232	0.202	14.452	0.000	2.527	3.320
Lead Source_Olark Chat	0.9922	0.119	8.303	0.000	0.758	1.226
Lead Source_Welingak Website	2.4324	0.746	3.261	0.001	0.970	3.895
Last Activity_SMS Sent	1.4351	0.073	19.552	0.000	1.291	1.579
Specialization_Select	-0.9756	0.122	-7.999	0.000	-1.215	-0.737
What is your current occupation_Working Professional	2.4208	0.191	12.674	0.000	2.046	2.795
What matters most to you in choosing a course_Other	-1.1468	0.087	-13.210	0.000	-1.317	-0.977
Last Notable Activity_Had a Phone Conversation	3.6198	1.126	3.214	0.001	1.412	5.827
Last Notable Activity_Unreachable	2.0552	0.551	3.733	0.000	0.976	3.134

Evaluation (Train set)

- According to the curve, 0.34 is the optimum point to take it as a cutoff probability
- Accuracy score: 0.81
- Sensitivity score: 0.81
- Specificity Score: 0.80





Evaluation (Test set)

- Accuracy score: 0.81
- Sensitivity score: 0.81
- Specificity Score: 0.80

Conclusion

- The company **should make calls** to the leads coming from the lead sources "Welingak Websites" and "Olark Chat" " as these are more likely to get converted.
- The company **should make calls** to the leads who are the "working professionals" as they are more likely to get converted.
- The company **should make calls** to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company **should make calls** to the leads whose last activity was ’SMS Sent’ as they are more likely to get converted.
- The company **should make calls** to the leads whose Last Notable Activity was ” Had a Phone Conversation” as they are more likely to get converted.

- The company **should not make calls** to the leads whose What matters most to you in choosing a course was "Other" as they are not likely to get converted.
- The company **should not make calls** to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company **should not make calls** to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company **should not make calls** to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.