Lead score case study

PRESENTED BY :

PADMASRI

VINEETH

CHIRAG

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.

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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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.

Goals of case study

- There are quite a few goals for this 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.

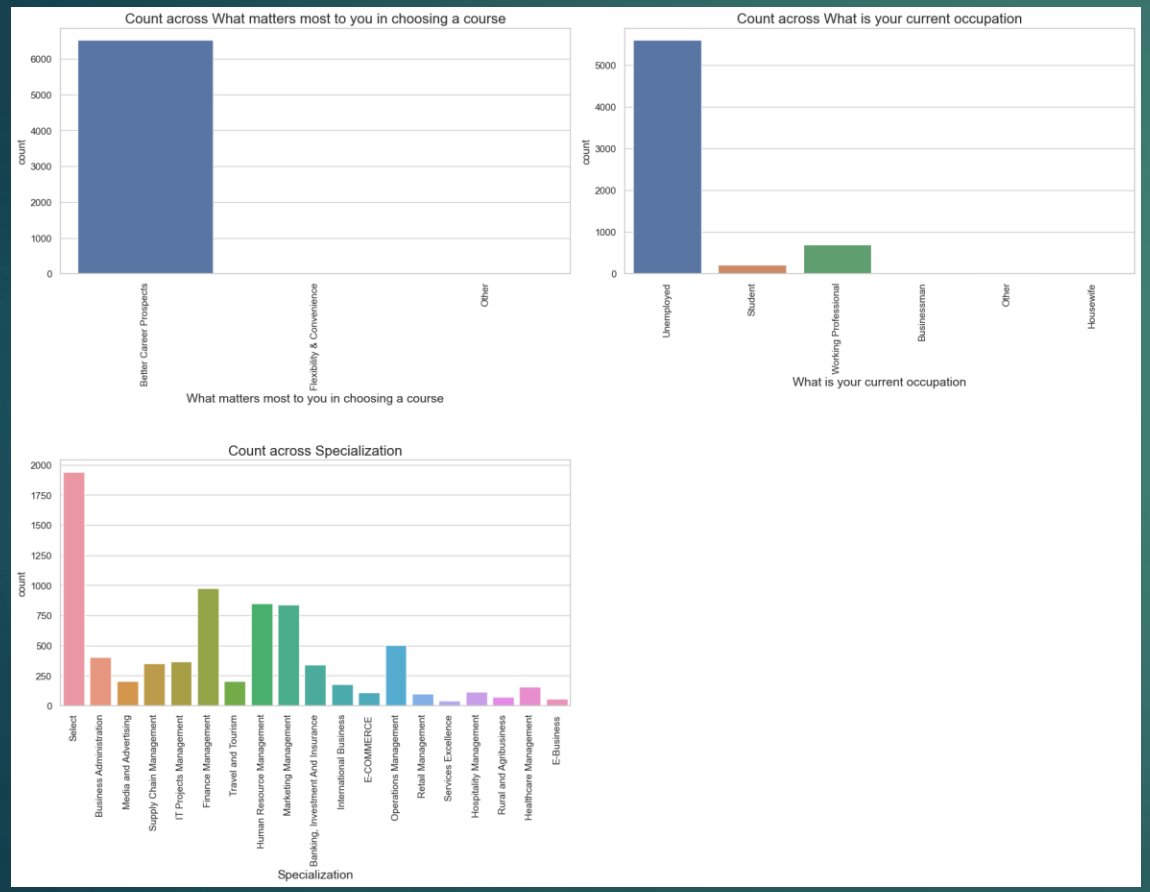
Approach

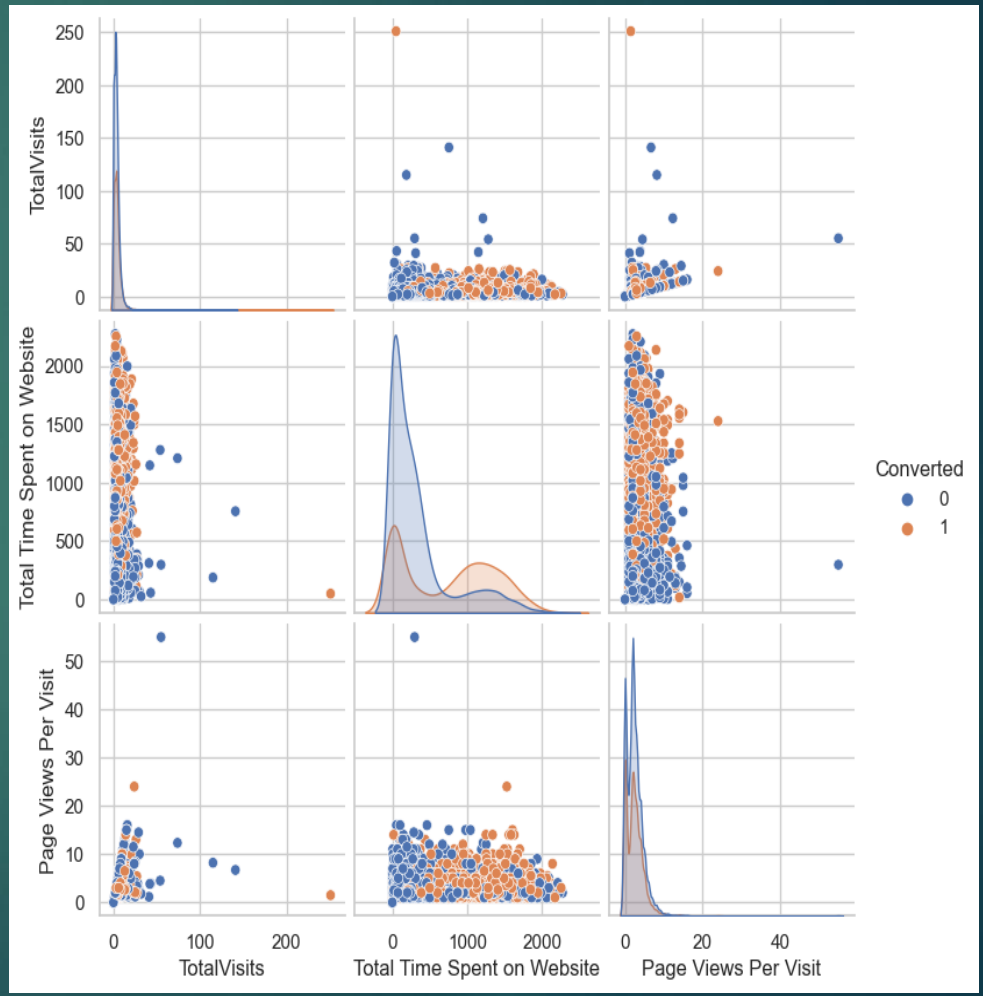
- Source the data For analysis
- Reading & Understanding the data
- Data Cleaning
- EDA Feature scaling
- Splitting the data into test & train dataset
- Prepare the data for modelling
- Model building
- Model evaluation-specificity & sensitivity or precision recall
- Making predictions on the test set

Data Sourcing, Cleaning & Preparation

- Read the data from CSV File
- Outlier treatment
- Data cleaning -Handling Null Values & removing higher Null values data
- Removing Redundant columns in the data. Imputing Null Values
- Exploratory data analysis-approx. Conversion Rate is 38%
- Feature standardization < 6 of 18 >

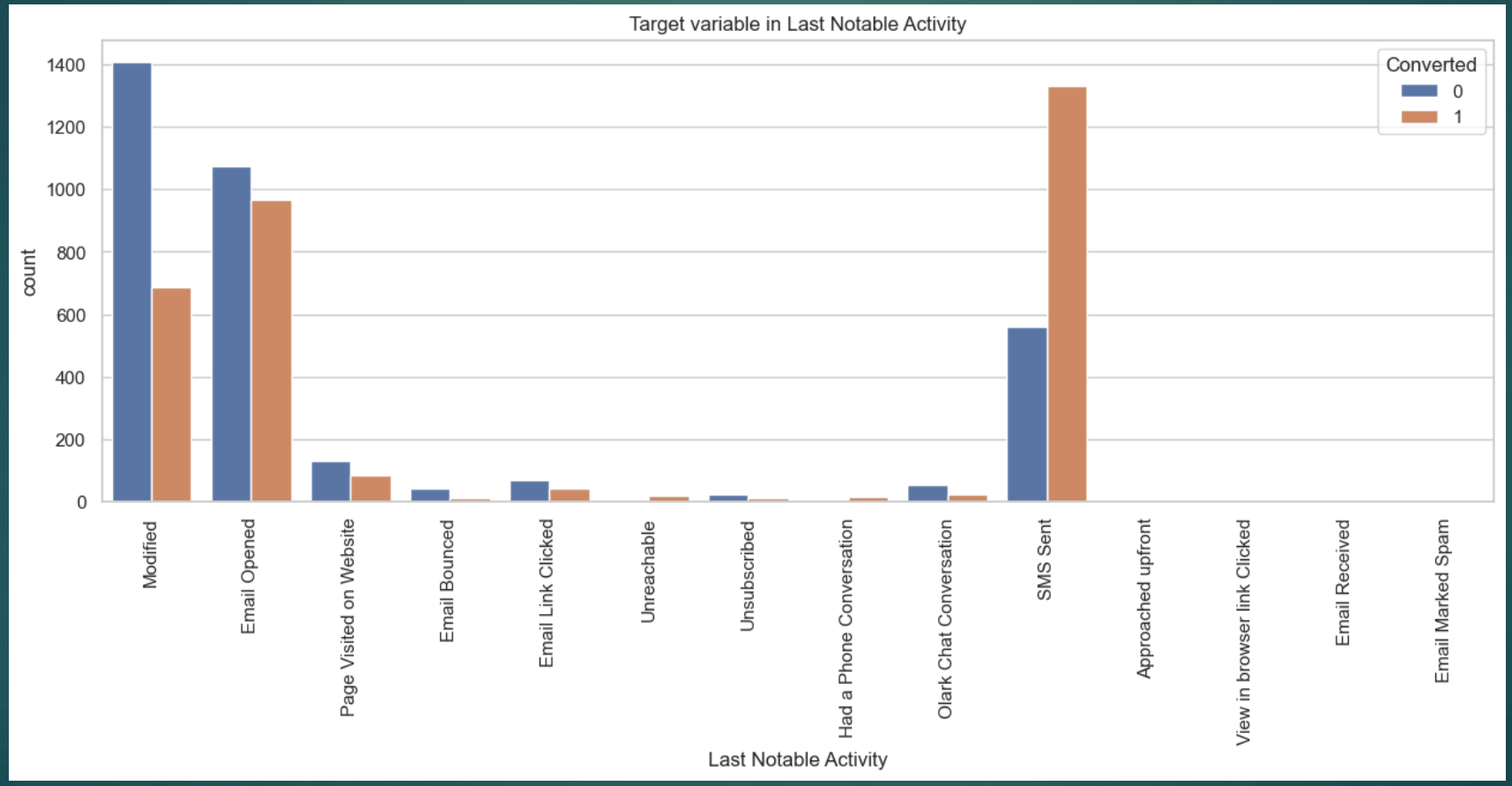
Visualizing the features



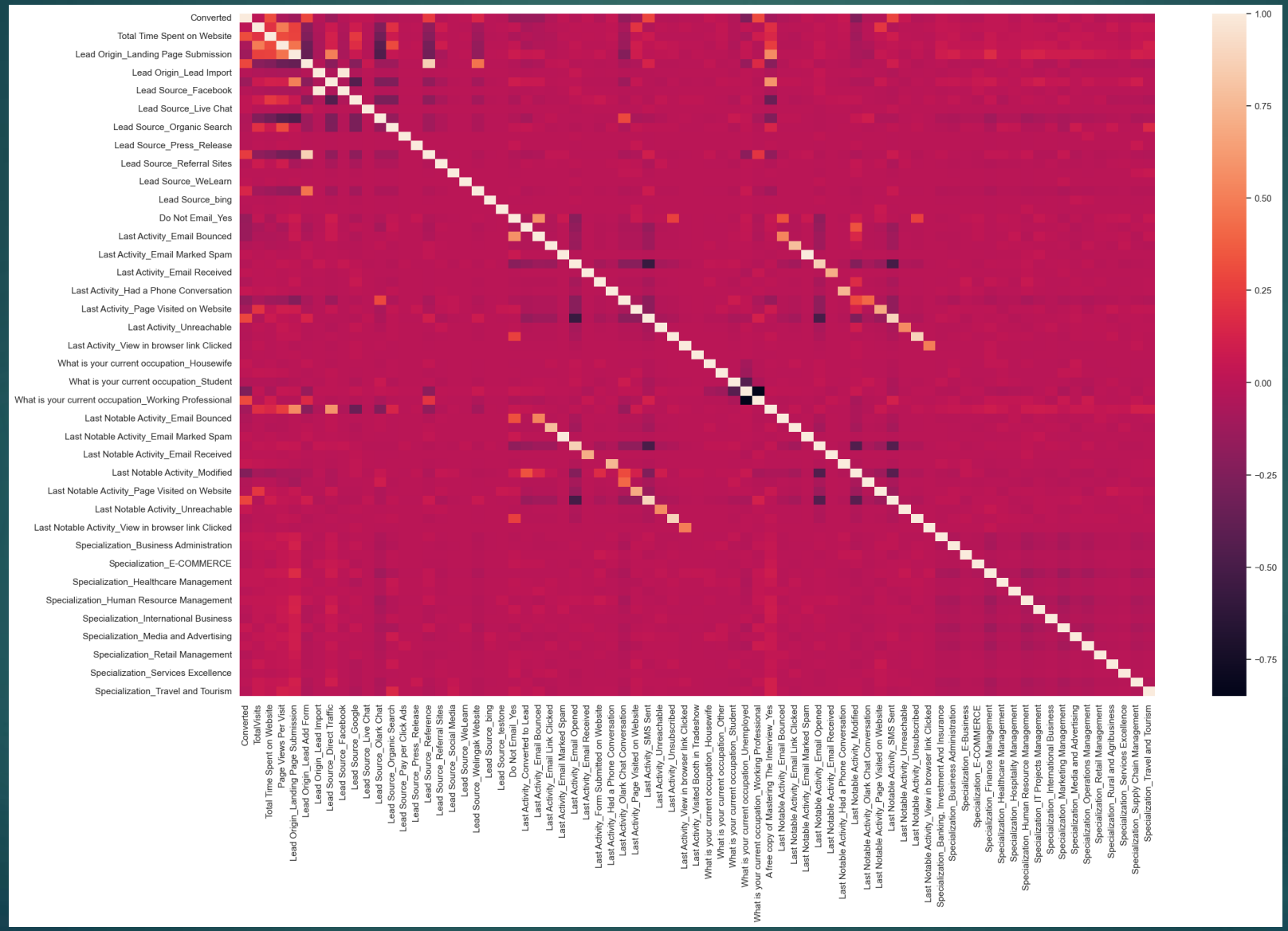



Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.098
0.005
0.039
0.035
0.098
1
0.0038
0.32
-0.066
0.005
0.0038
1
0.2
0.49
0.039
0.32
0.2
1
0.31
0.035
-0.066
0.49
0.31
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Lead Number
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit

Analysing Categorical features



Looking at the correlations



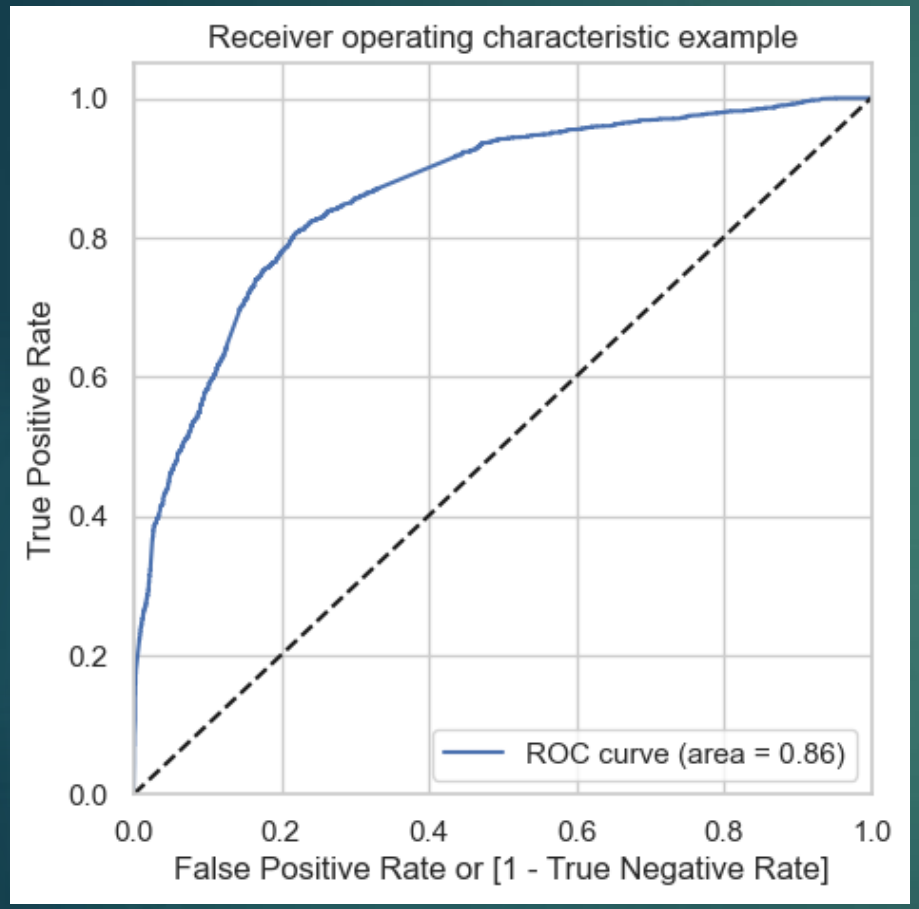
Data preparation

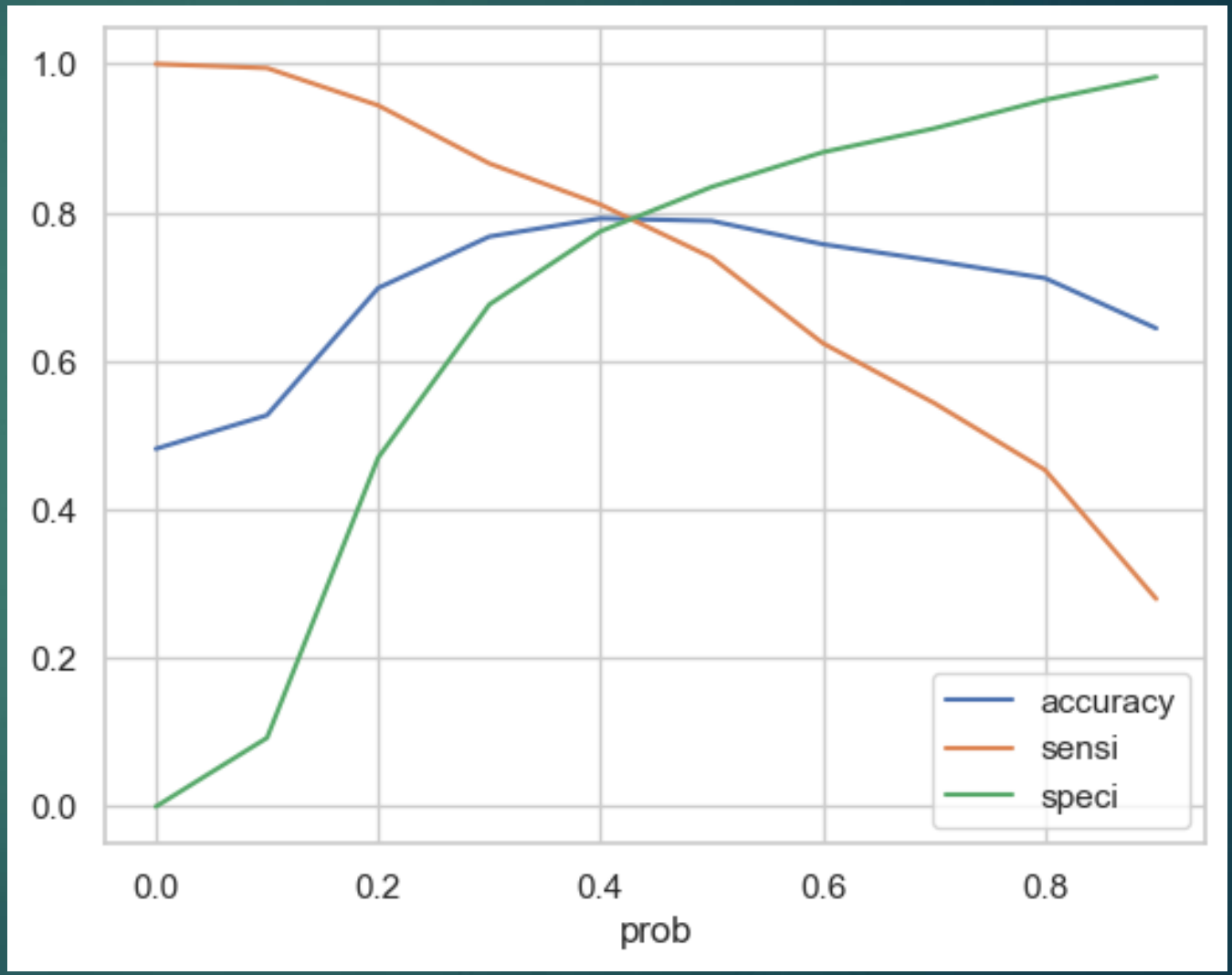
- **Converted binary variable into 0to 1.**
- **Created dummy variables for categorical variables.**

Model Building

- Feature selection using RFE
- Determined optimal model using logistic regression
- Caluculated accuracy, sensitivity,specificity , precision and recall and evaluate model

Finding the Optimal Cutoff





Model building

```
# Fit a logistic Regression model on X_train after adding a constant and output the sum

X_train_sm = sm.add_constant(X_train)
logm2 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4445
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2072.8
Date:	Wed, 16 Aug 2023	**Deviance:**	4145.5
Time:	16:15:44	**Pearson chi2:**	4.84e+03
No. Iterations:	22	**Pseudo R-squ. (CS):**	0.3660
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1932	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7868	0.807	3.453	0.001	1.205	4.369

Model 2

```
# Refit the model with the new set of features

logm1 = sm.GLM(y_train,(sm.add_constant(X_train)), family = sm.families.Binomial())
logm1.fit().summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4446
Model Family:	Binomial	**Df Model:**	14
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2073.2
Date:	Wed, 16 Aug 2023	**Deviance:**	4146.5
Time:	16:16:07	**Pearson chi2:**	4.82e+03
No. Iterations:	22	**Pseudo R-squ. (CS):**	0.3658
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0057	0.600	-1.677	0.094	-2.181	0.170
TotalVisits	11.3428	2.682	4.229	0.000	6.086	16.599
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	4.2084	0.259	16.277	0.000	3.702	4.715
Lead Source_Olark Chat	1.4583	0.122	11.960	0.000	1.219	1.697
Lead Source_Welingak Website	2.1557	1.037	2.079	0.038	0.124	4.188
Do Not Email_Yes	-1.5036	0.193	-7.779	0.000	-1.882	-1.125
Last Activity_Had a Phone Conversation	1.0398	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6511	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1537	0.630	-1.830	0.067	-2.389	0.082
What is your current occupation_Unemployed	-1.3401	0.594	-2.255	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2748	0.623	2.046	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1934	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7872	0.807	3.454	0.001	1.205	4.369

Model 3

```
# Refit the model with the new set of features

logm1 = sm.GLM(y_train,(sm.add_constant(X_train)), family = sm.families.Binomial())
logm1.fit().summary()
```

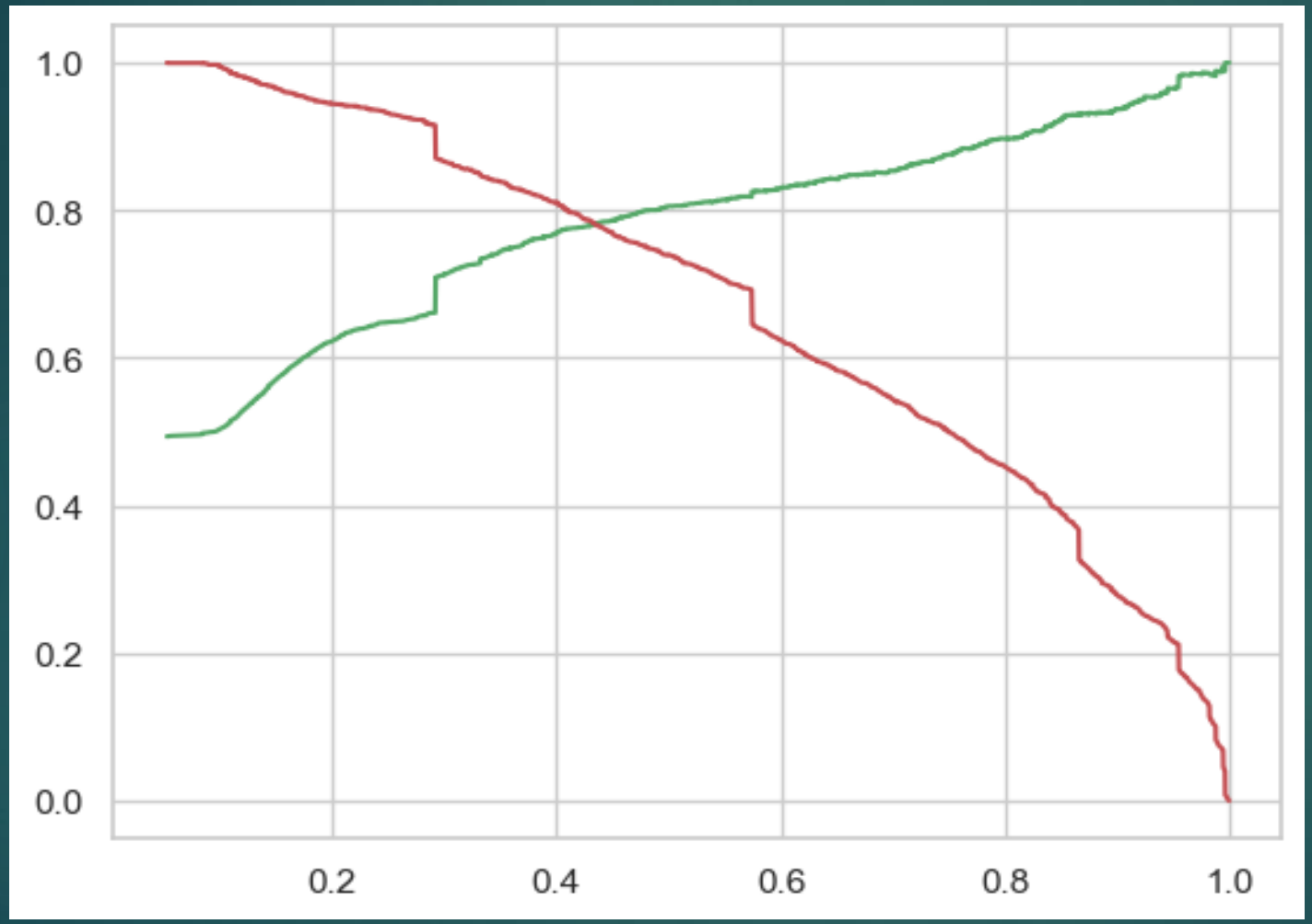
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4447
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2076.1
Date:	Wed, 16 Aug 2023	**Deviance:**	4152.2
Time:	16:16:20	**Pearson chi2:**	4.82e+03
No. Iterations:	21	**Pseudo R-squ. (CS):**	0.3650
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0069	0.600	-1.679	0.093	-2.182	0.168
TotalVisits	11.4551	2.686	4.265	0.000	6.191	16.720
Total Time Spent on Website	4.4237	0.185	23.900	0.000	4.061	4.787
Lead Origin_Lead Add Form	4.2082	0.259	16.276	0.000	3.701	4.715
Lead Source_Olark Chat	1.4581	0.122	11.958	0.000	1.219	1.697
Lead Source_Welingak Website	2.1557	1.037	2.079	0.038	0.124	4.188
Do Not Email_Yes	-1.5037	0.193	-7.780	0.000	-1.882	-1.125
Last Activity_Had a Phone Conversation	2.7502	0.802	3.430	0.001	1.179	4.322
Last Activity_SMS Sent	1.1826	0.082	14.364	0.000	1.021	1.344
What is your current occupation_Housewife	21.6525	1.49e+04	0.001	0.999	-2.91e+04	2.91e+04
What is your current occupation_Student	-1.1520	0.630	-1.828	0.068	-2.387	0.083
What is your current occupation_Unemployed	-1.3385	0.594	-2.253	0.024	-2.503	-0.174
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.495
Last Notable Activity_Unreachable	2.7862	0.807	3.453	0.001	1.205	4.368

Dropping the `What is your current occupation_Housewife` as having high P value

Precision and recall tradeoff



Result

- Accuracy, Sensitivity and Specificity values of training and test set are close to training set ► Accuracy, Sensitivity and Specificity values of training set are79%,82%,76% Respectively ► Accuracy, sensitivity & Specificity values of test are 78%,81%,76% Respectively Conversion rate for Train & Test Dataset Is 82.7% & 80.8% Respectively We have done the prediction on the test set using cut off threshold from sensitivity & specificity metrics 16

Conclusion

- While we have checked both sensitivity-specificity as well as Precision & recall metrics, we have considered the optimal cut off based on sensitivity & specificity for calculating the final prediction Accuracy, Sensitivity & specificity values of test set are around 78%,81%,76% which are approximately closer to Values calculated using Trained Data Set Lead Score Calculated for the conversion rate final model on Train & Test dataset is 82.7% &80.8% respectively.
- Hence, Overall Model seems to be Good

Summery

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversionFirst, sort out the best prospects from the leads you have generated. 'Total Visits, 'Total Time Spent on Website', 'Page Views Per Visit' which contribute most towards the probability of a lead getting convertedThen, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to themCarefully provide job offerings, information or courses that suits best according to the interest of the leadsA proper plan to chart the needs of each lead will go a long way to capture the leads as prospects. Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about themMake further inquiries and appointments with the leads to determine their intention and mentality to join online courses.