

Lead Scoring Case Study

Submitted By

Fasna C K

Ratheesh Kumar

Contents

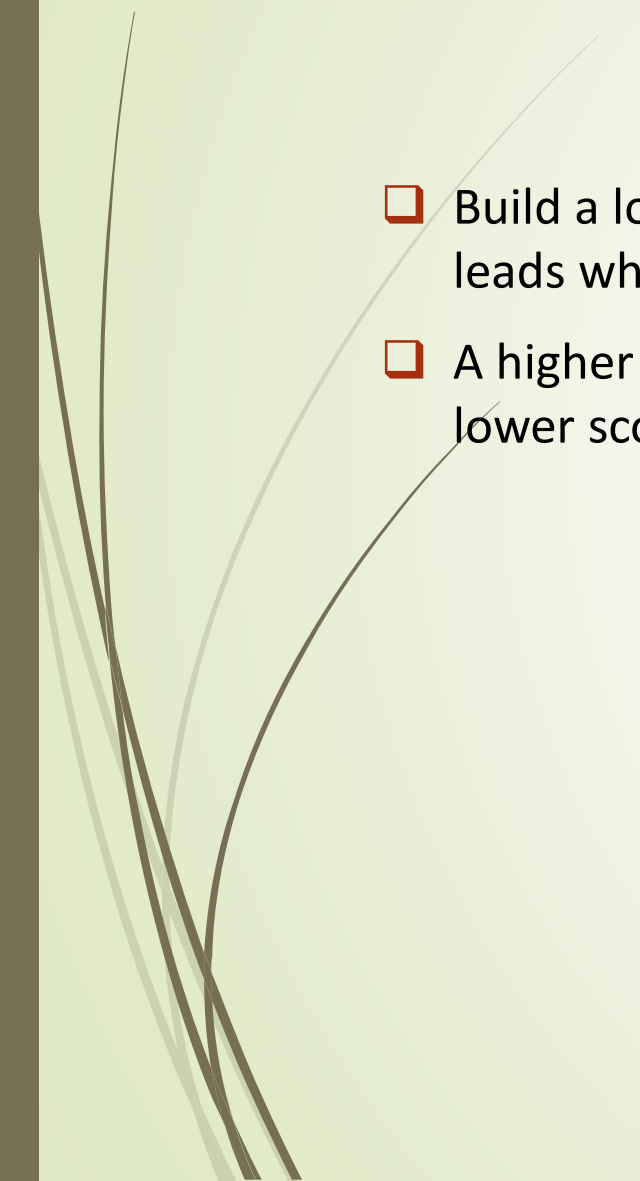
- Problem Statement
- Goals of Case Study
- Problem Approach
- EDA
- Correlations
- Model Building
- Model Evaluation
- Observations
- Conclusion

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. They have process of form filling on their website after which the company that individual as 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%. Now, this means 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

- ☐ 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.
- 

Problem Approach

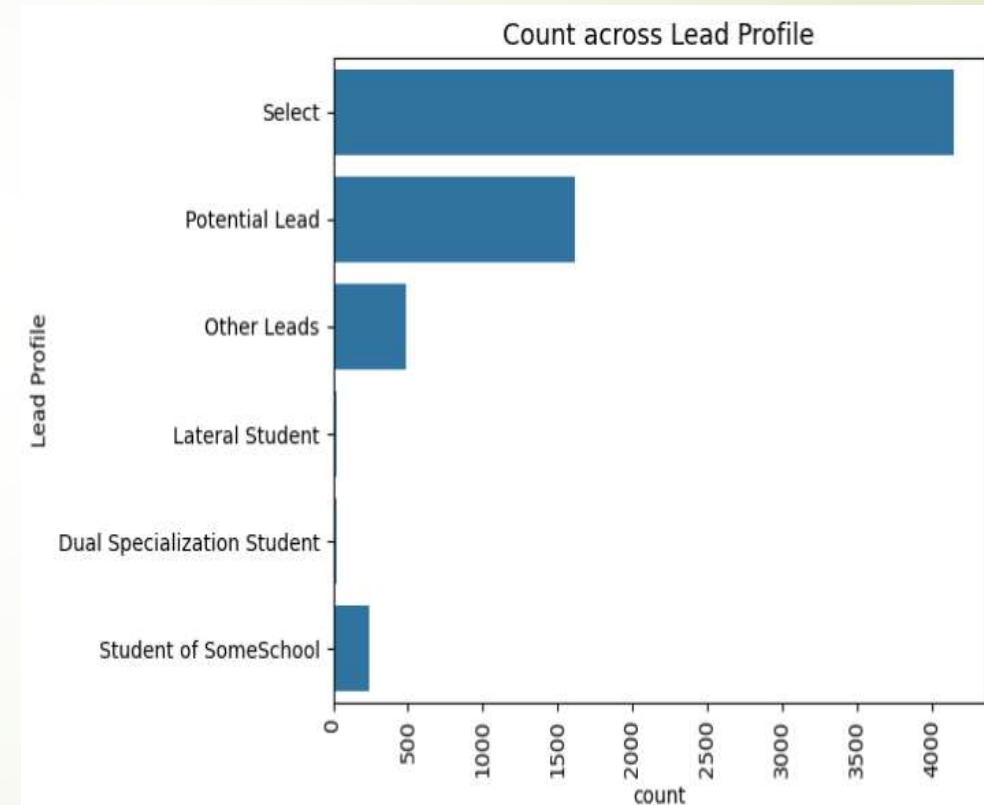
1. Importing the data and inspecting the data frame
2. Data cleaning and preparation
3. EDA
4. Dummy variable creation
5. Test-Train split
6. Feature scaling
7. Correlations
8. Model Building
9. Model Evaluation
10. Prediction on test set
11. Lead score assigning

EDA – Data Cleaning

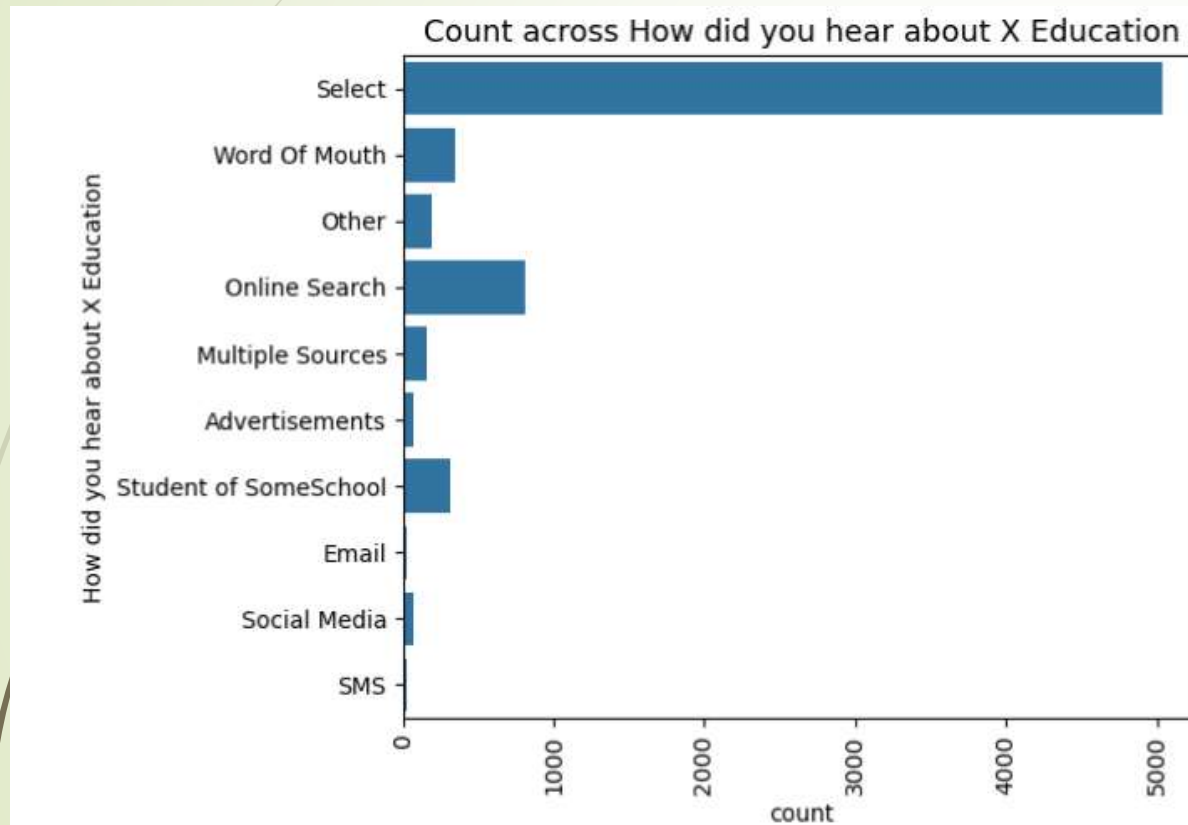
There are a few columns in which there is a level called 'Select' which is taking care

Feature 'Lead Profile'

- Feature 'Lead Profile' has 4146 'Select' data points.
- Because of higher number of 'Select' data points, which are not useful for data analysis, feature 'Lead Profile' is dropped

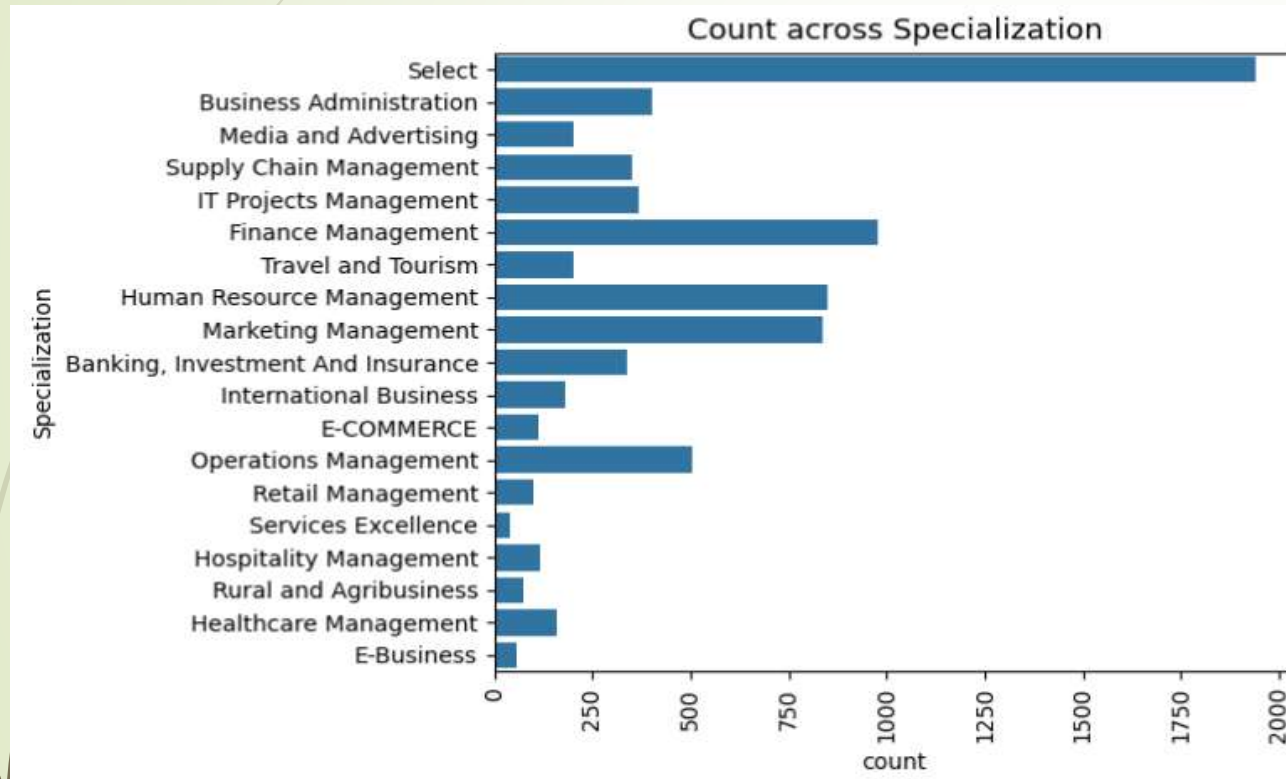


Feature 'How did you hear about X Education'



- Feature 'How did you hear about X Education' has 5043 'Select' data points.
- Because of higher number of 'Select' data points, which are not useful for data analysis, feature 'How did you hear about X Education' is dropped from the data set.

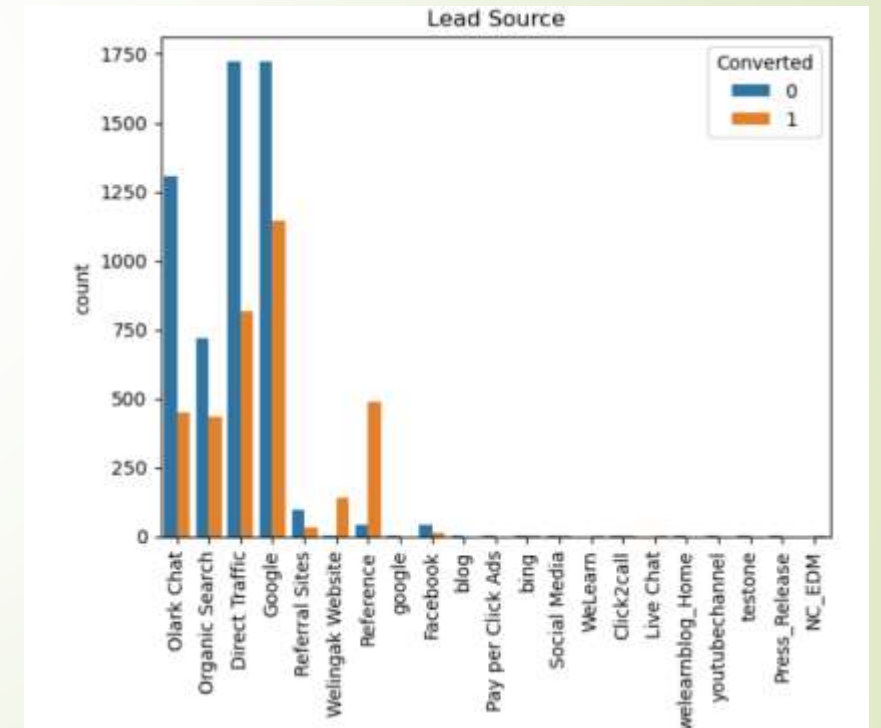
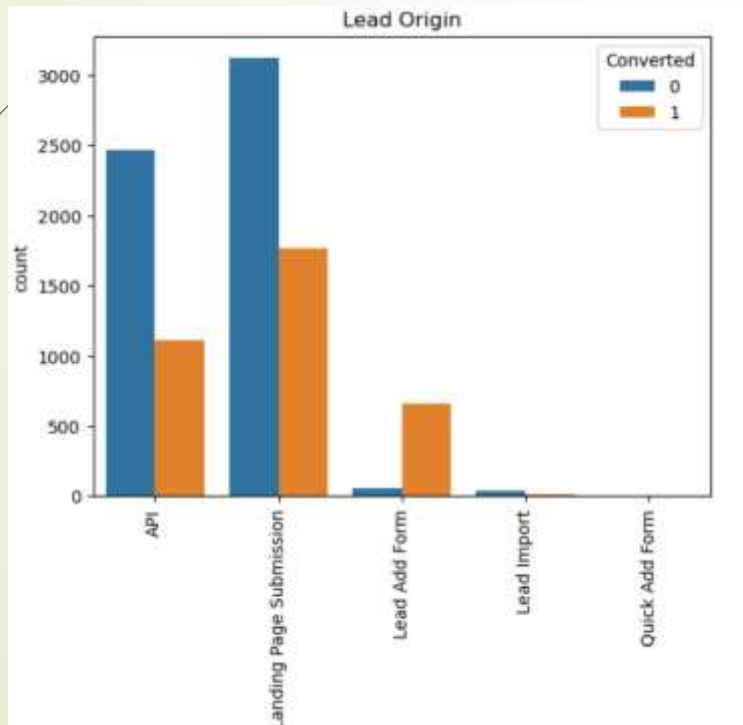
Feature 'Specialization'



- Feature 'Specialization' has 1942 'Select' data points.
- Because of less number of 'Select' data points, which are must useful for data analysis, feature "How did you hear about X Education" is not dropped.
- 'Select' labels are dropped during dummy creation.
- Leads from HR, Finance & Marketing management specializations are high probability to convert.

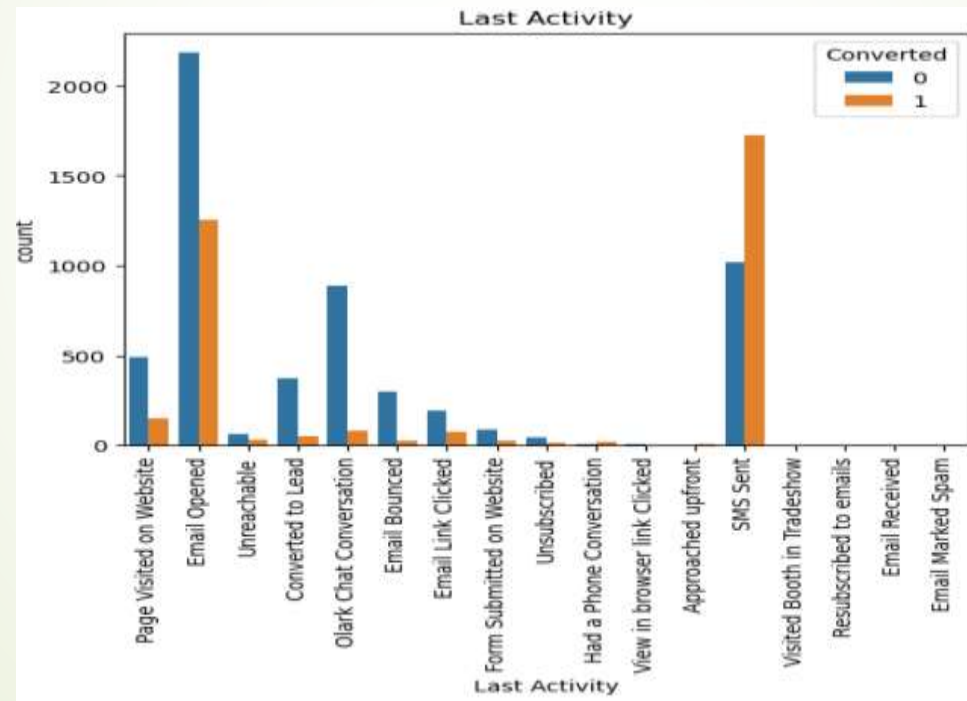
Lead Source & Lead Origin

- In lead source the leads through google & direct traffic high probability of lead conversion
- In Lead origin a greater number of leads are landing on submission.

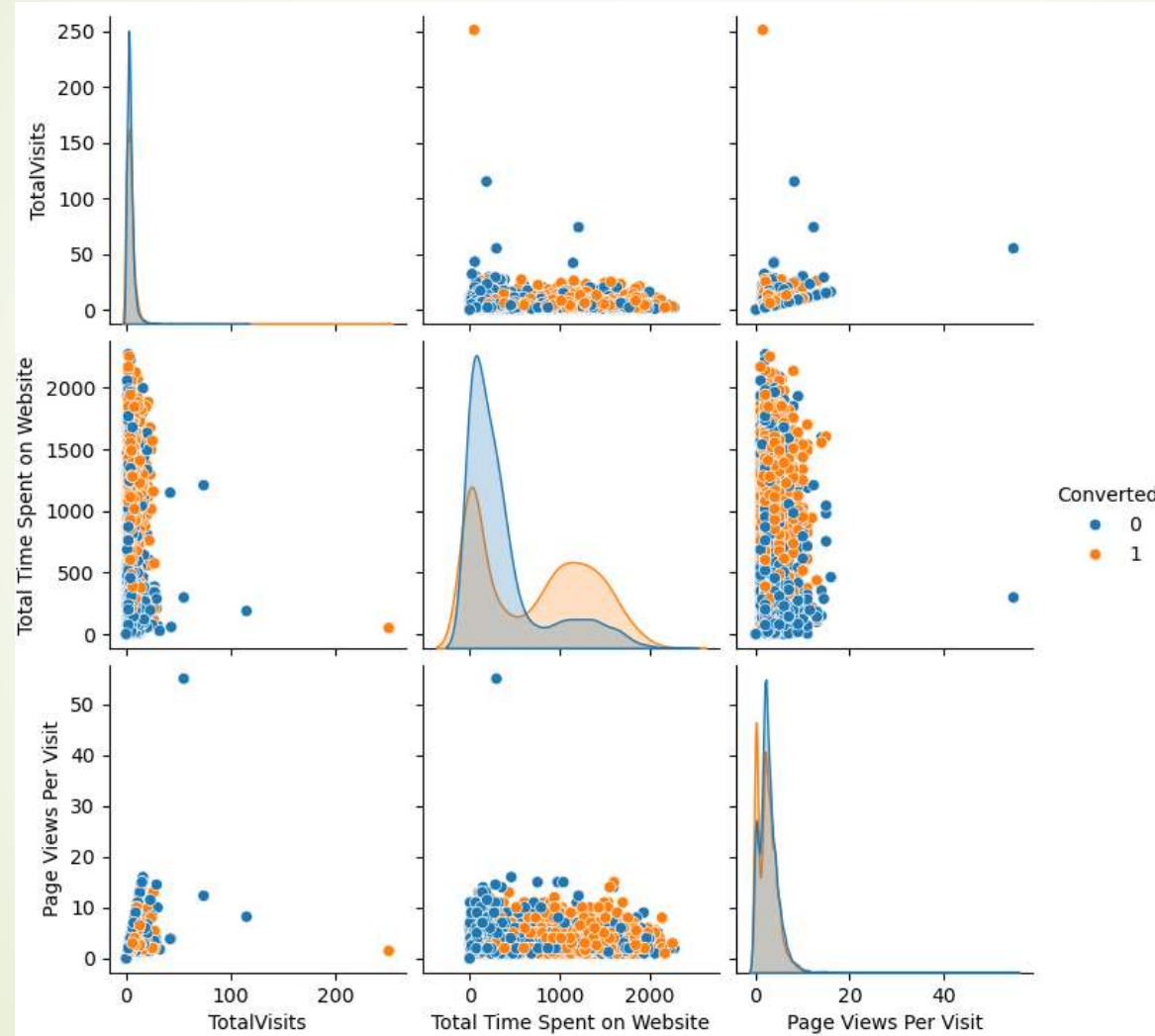


Last Activity

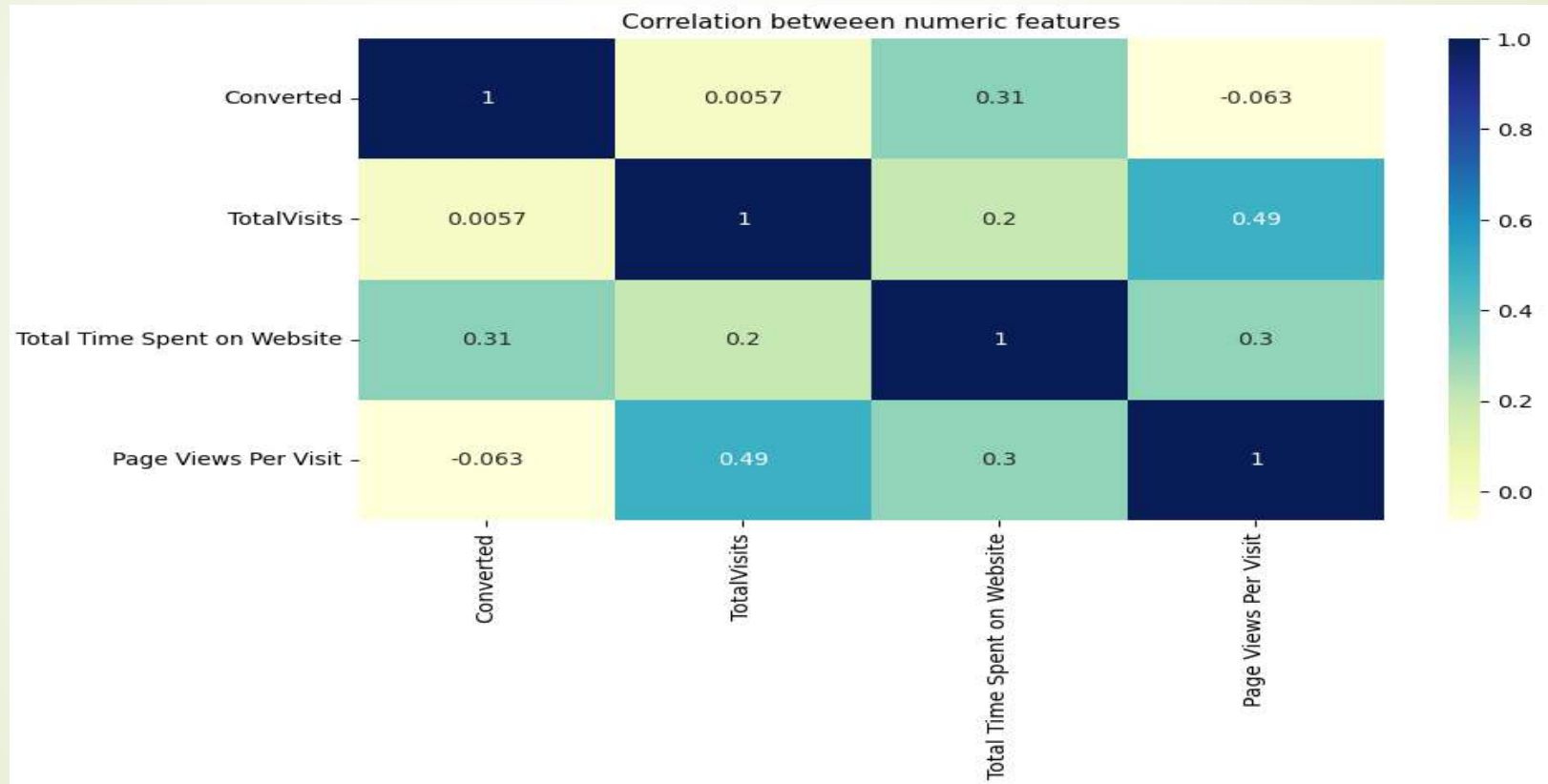
Leads which are opening email have high probability to convert, Same as Sending SMS will also benefit.



Pair Plot of Numerical Columns




Correlation Between Numerical Features





Train-Test Split

- Target Variable (y)- Converted
 - Independent Variable(X)-Total 74 variables.
 - Splitting Train data -70%
 - Splitting Test data-30%
 - Assume Random State-100
 - Feature Scaling- Minmax Scaler
- 



Model Building

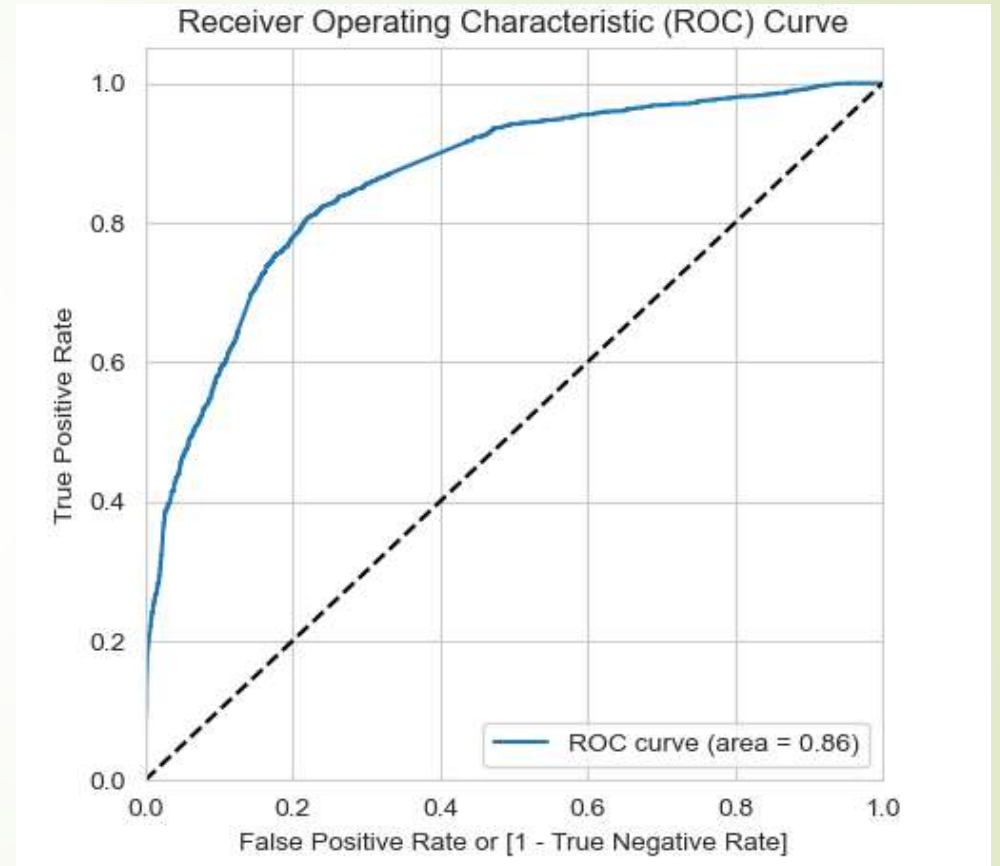
- Feature Selection – RFE a logistic regression model was built in Python using the function GLM() under statsmodel library.
- Some variables were removed first based on an automated approach, i.e. RFE (Running RFE with 15 variables)
- Manual approach based on the VIFs and p-values are used for further process.
- After dropping some columns which has p-values above 0.05 and VIF greater than 5, we have obtained our final model with 11 variables.

Features For Final Model With VIF

Features	VIF
What is your current occupation_Unemployed	2.82
Total Time Spent on Website	2.00
TotalVisits	1.54
Last Activity_SMS Sent	1.51
Lead Origin_LeadAdd Form	1.45
Lead Source_Olark Chat	1.33
Lead Source_Welingak Website	1.30
Do Not Email	1.08
What is your current occupation_Student	1.06
Last Activity_Had a Phone Conversation	1.01
Last Notable Activity_Unreachable	1.01

ROC Curve

- It shows the tradeoff between sensitivity and specificity
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The area under the curve of the ROC is
- 0.86 which is quite good.



Metrics for Different Probability Cutoff

Probability Cutoff	Accuracy	Sensitivity	Specificity
0.0	0.481731	1.000000	0.000000
0.1	0.527012	0.994416	0.092561
0.2	0.698274	0.944160	0.469723
0.3	0.767541	0.865984	0.676038
0.4	0.791975	0.810610	0.774654
0.5	0.788612	0.739414	0.834343
0.6	0.757229	0.624011	0.881055
0.7	0.735037	0.543509	0.913062
0.8	0.711500	0.453234	0.951557
0.9	0.644026	0.279665	0.982699

Finding Optimal Cutoff

We have checked optimal cutoff using Sensitivity Specificity and accuracy and also using another pair of industry-relevant metric precision and recall Tradeoff .

The optimum point has cutoff probability is 0.42.

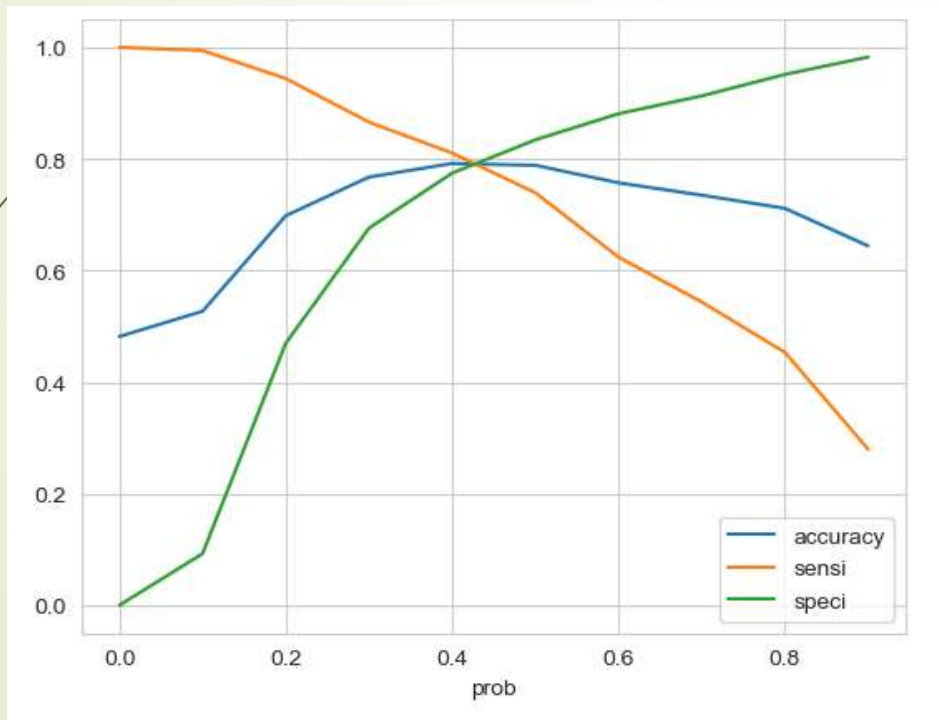


Fig a: Precision Recall Tradeoff Curve

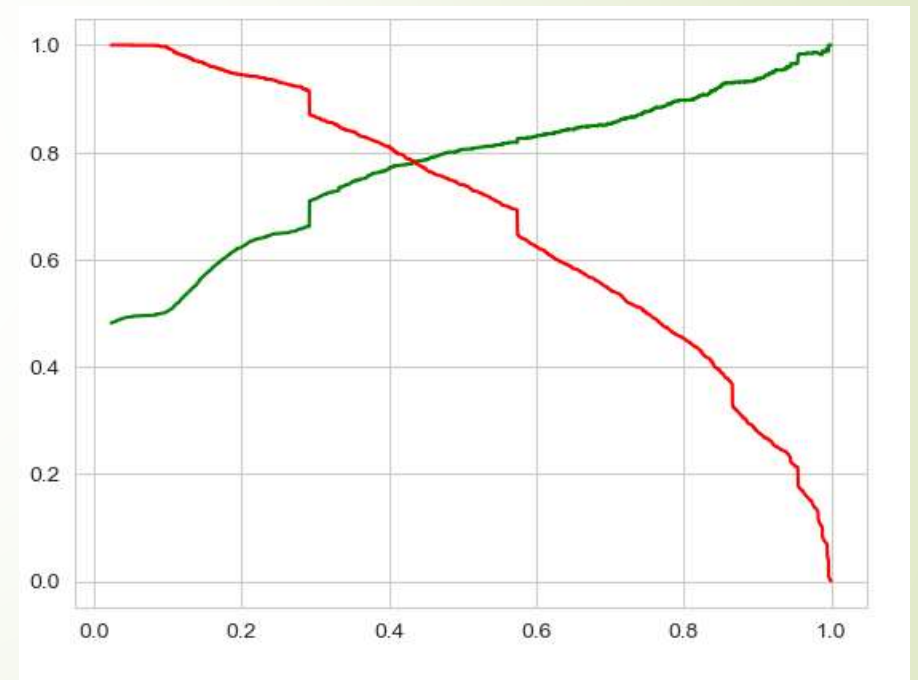


Fig b: Cutoff using Sensitivity Specificity and Accuracy

Confusion Matrix

Trained Data

Actual/ Predicted	Not Converted	Converted
Not Converted	1823	489
Converted	444	1705

Test Data

Actual/ Predicted	Not Converted	Converted
Not Converted	786	210
Converted	202	714

Making Predictions on Test Data

- After checking precision and recall tradeoff curve we have chosen a cut-off point of 0.42.
- Made predictions on the test dataset and have obtained a descent values for Accuracy (78.45%),Sensitivity (77.94%),Specificity (78.92%).



Result for Trained Data

1. Accuracy 79.08%
2. Sensitivity 79.34%
3. Specificity 78.85%
4. Precision 77.71%
5. Recall 79.34%

Result for Test Data

1. Accuracy 78.45%
2. Sensitivity 77.94%
3. Specificity 78.92%
4. Precision 77.27%
5. Recall 77.95%

Finally selected Features with Coefficients

	coef
const	0.2040
TotalVisits	11.1489
Total Time Spent on Website	4.4223
Lead Origin_Lead Add Form	4.2051
Lead Source_Olark Chat	1.4526
Lead Source_Welingak Website	2.1526
Do Not Email_Yes	-1.5037
Last Activity_Had a Phone Conversation	2.7552
Last Activity_SMS Sent	1.1856
What is your current occupation_Student	-2.3578
What is your current occupation_Unemployed	-2.5445
Last Notable Activity_Unreachable	2.7846

Conclusion

- According to the model 932 hot leads are predicted as 1, it shows they have high chance of getting converted.
- **Company should concentrate more on following features :**
 1. TotalVisits – Leads who visited more in Xeducation site.
 2. Total Time Spent on Website – Leads who spent more time on Xeducation Website
 3. Lead Origin_Lead Add Form – Lead Origin category where Lead Add form
 4. Last Activity_Had a Phone Conversation and Last Activity_SMS Sent – Lead who had a phone conversation and sent SMS have high chance of getting converted.
 5. Lead Source_Wellingak Website and Lead Source_Olark Chat – Source of the lead from Wellingak Website and Olark Chat have high conversion rate.