X Education - Report

Data Preparation

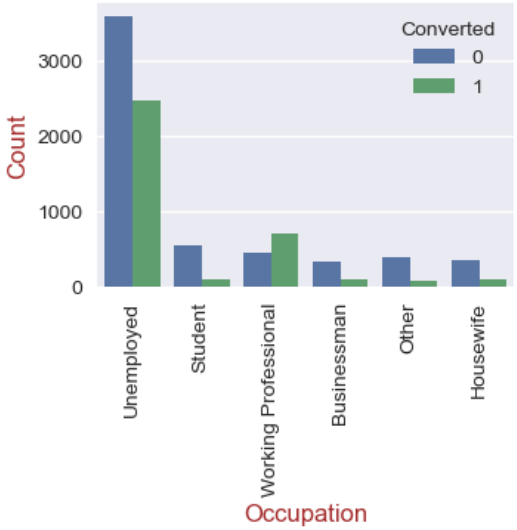
- Dataset originally contains 37 columns and 9240 records.
- Following columns has been removed because these are unique numbers and index:
    - Prospect ID
    - Lead Number
- Following columns have been removed because these are not the key drivers for the analysis:
    - Update me on Supply Chain Content
    - Get updates on DM Content
    - I agree to pay the amount through cheque.
    - Tags
- Certain column names are given aliases.
- Target variable (Y) is Converted
- There are few values called Select, will convert those to NAN and will treat with other missing values.

Impute/Remove Missing Values

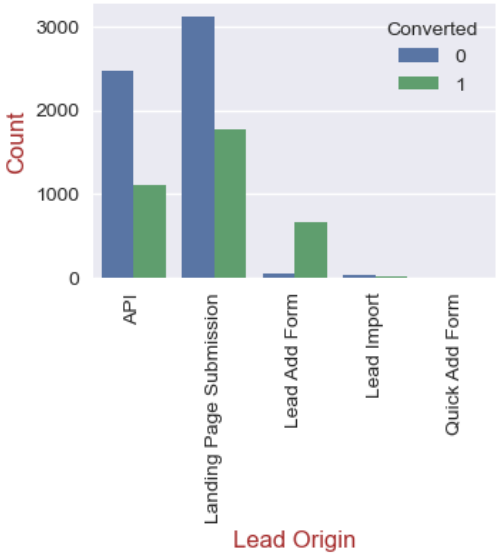
- Total 16 columns are having null values.
- Removed all the null values above 40%
- Removed following columns because after null value imputation high skewness in the data was found.
    - Country
    - Moto alias of "What matters most to you in choosing a course".
    - City
- Distributed null values uniformly for following columns because to maintain uniformity and avoid skewness:
    - Specialization
    - Occupation
    - Lead Source
    - Last Activity
- Imputed null values for following columns with null values:
    - Total Visits
    - Page Views Per Visit

Data Visualization

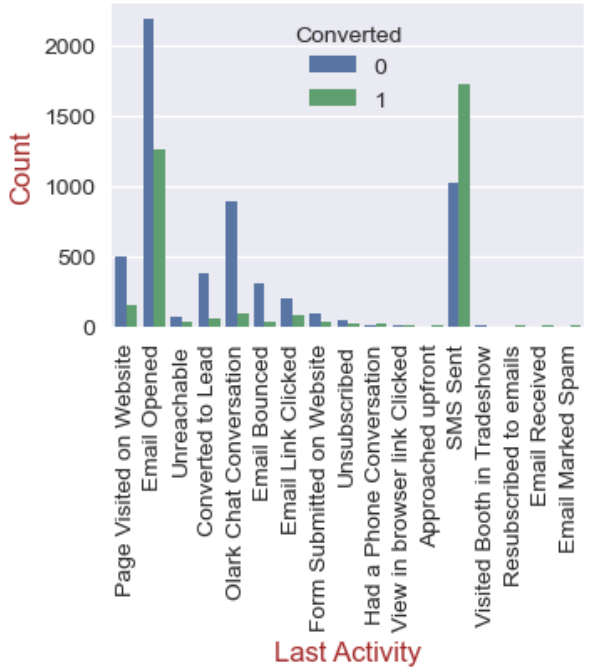
Occupation: The below chart shows that the working professionals are more interested in the courses.



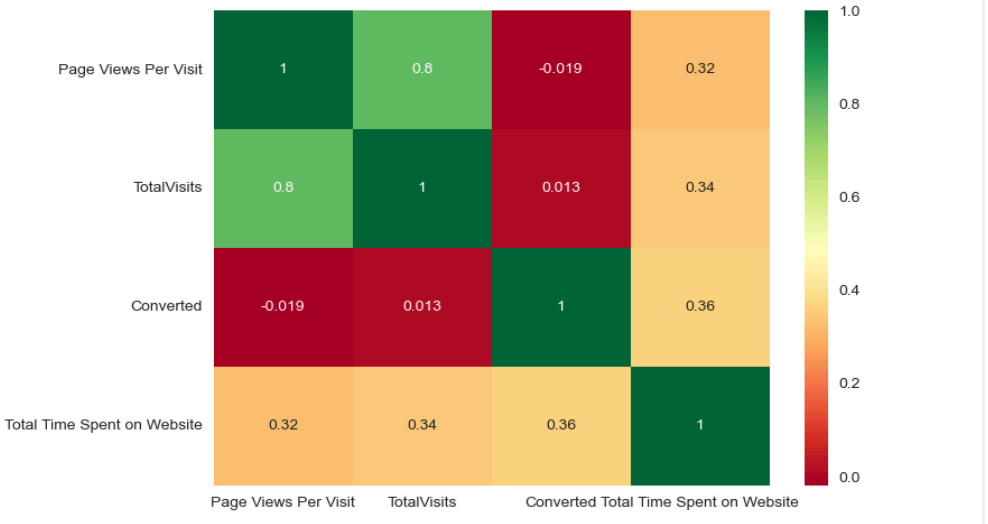
Lead Origin: The below chart shows that the customers who are filling the lead form are getting converted more.



Last Activity: The conversion rate is high for the customers to whom the “SMS sent”, shows the below chart.



Total Time Spent on Website: The total visits and total time the user spends on website shows higher corelation with the conversion rate.



Model Training

Model 1: Created using all the columns.

Conclusion: P value and VIF is quite high for many features. Hence better to perform the feature selection using RFE.

Model 2: Model created by features selected by RFE.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6447
Model Family:	Binomial	Df Model:	20
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2824.1
Date:	Sat, 13 Jan 2024	Deviance:	5648.2
Time:	23:27:41	Pearson chi2:	6.80e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3663
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8709	0.205	-13.994	0.000	-3.273	-2.469
Do Not Email	-0.3254	0.048	-6.839	0.000	-0.419	-0.232
Total Time Spent on Website	1.1039	0.039	28.566	0.000	1.028	1.180
Lead Add Form	3.6758	0.282	13.037	0.000	3.123	4.228
Lead Import	-0.4551	0.560	-0.813	0.416	-1.552	0.642
LeadSource_Others	0.2025	0.241	0.841	0.400	-0.269	0.674
Olark Chat	1.1513	0.101	11.407	0.000	0.953	1.349
Email Link Clicked	0.5924	0.271	2.187	0.029	0.061	1.123
Email Opened	1.2036	0.181	6.639	0.000	0.848	1.559
Form Submitted on Website	0.5929	0.356	1.665	0.096	-0.105	1.291
LastActivity_Others	2.2130	0.342	6.478	0.000	1.543	2.883
Olark Chat Conversation	-0.3120	0.234	-1.331	0.183	-0.771	0.147
Page Visited on Website	0.7005	0.214	3.274	0.001	0.281	1.120
SMS Sent	2.3116	0.181	12.770	0.000	1.957	2.666
Unreachable	1.2868	0.381	3.377	0.001	0.540	2.033
Other	-0.4432	0.209	-2.124	0.034	-0.852	-0.034
Student	-0.5424	0.190	-2.858	0.004	-0.914	-0.170
Unemployed	0.4813	0.120	4.022	0.000	0.247	0.716
Working Professional	1.3754	0.146	9.440	0.000	1.090	1.661
Hospitality Management	-0.8698	0.304	-2.864	0.004	-1.465	-0.275
Rural and Agribusiness	0.6984	0.378	1.846	0.065	-0.043	1.440

Conclusion: P-value is high for few features, that can be treated manually.

Model 3:

Create model after dropping below features having high p value:

- Lead import
- LeadSource_Others
- Form Submitted on Website
- Olark Chat Conversion
- Rural and agribusiness

Generalized Linear Model Regression Results

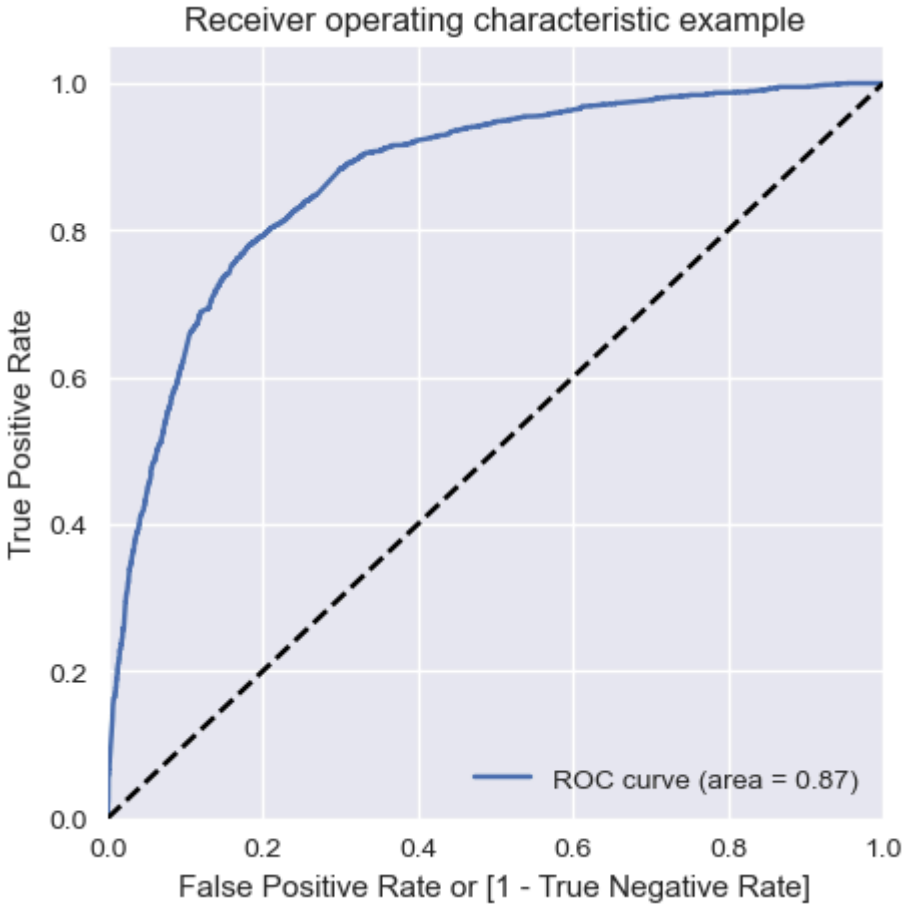
Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2829.6
Date:	Sat, 13 Jan 2024	Deviance:	5659.2
Time:	23:27:41	Pearson chi2:	6.81e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3652
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9573	0.157	-18.803	0.000	-3.266	-2.649
Do Not Email	-0.3180	0.047	-6.789	0.000	-0.410	-0.226
Total Time Spent on Website	1.1063	0.039	28.678	0.000	1.031	1.182
Lead Add Form	3.8612	0.179	21.583	0.000	3.511	4.212
Olark Chat	1.1039	0.098	11.252	0.000	0.912	1.296
Email Link Clicked	0.7013	0.232	3.023	0.003	0.247	1.156
Email Opened	1.3044	0.118	11.026	0.000	1.073	1.536
LastActivity_Others	2.3116	0.321	7.192	0.000	1.682	2.942
Page Visited on Website	0.7979	0.167	4.770	0.000	0.470	1.126
SMS Sent	2.4053	0.121	19.915	0.000	2.169	2.642
Unreachable	1.3773	0.358	3.852	0.000	0.676	2.078
Other	-0.4527	0.208	-2.174	0.030	-0.861	-0.045
Student	-0.5327	0.189	-2.812	0.005	-0.904	-0.161
Unemployed	0.4881	0.119	4.088	0.000	0.254	0.722
Working Professional	1.3787	0.145	9.490	0.000	1.094	1.663
Hospitality Management	-0.8815	0.304	-2.902	0.004	-1.477	-0.286

Conclusion: Model 3 looks good because after removing the column, the results show:

- p value < 0.05, and
- VIF < 10

ROC

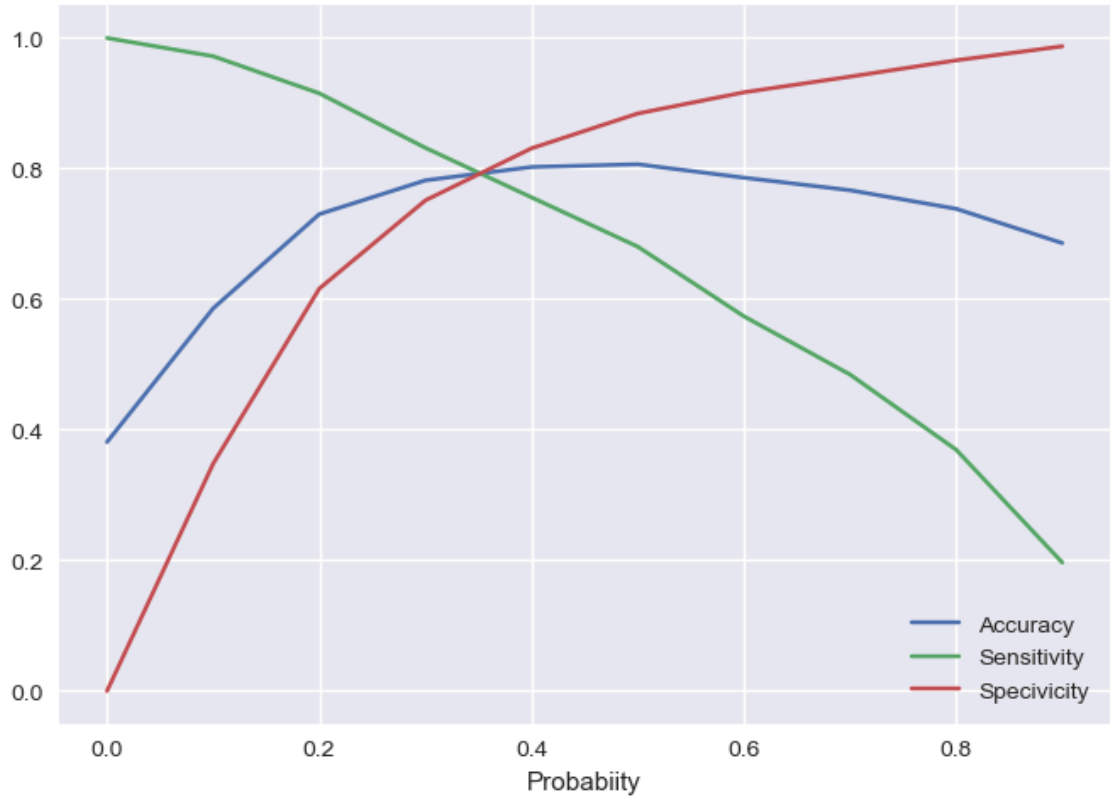


Area of ROC is 0.87 which is a good curve:

- Accuracy: 80.50
- Sensitivity: 68.00
- Specificity: 88.205
- False Positive Rate: 11.79
- Precision: 78.03

Above values are from the randomly selected threshold = 0.5. Since the Sensitivity is very less, the threshold needs to be optimal.

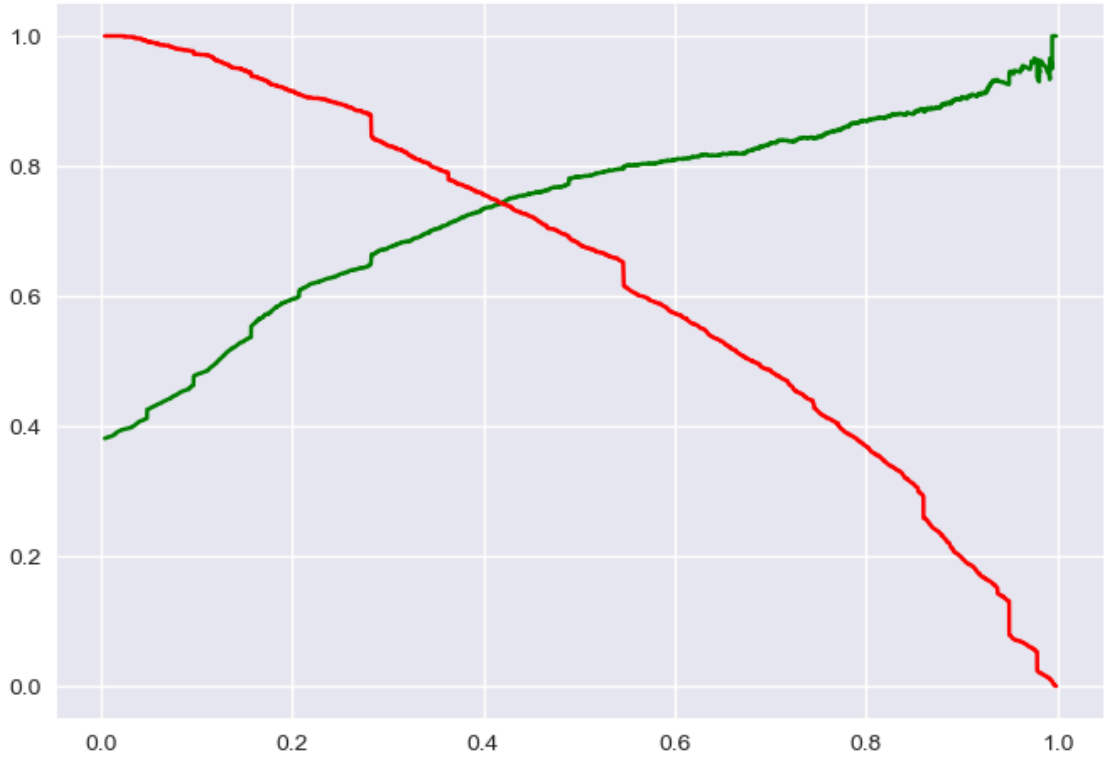
Optimum Threshold Calculation



From the above curve, 0.352 is the optimum point to take it as a cutoff probability.

- Accuracy: 80.24
- Sensitivity: 78.02
- Specificity: 81.60
- False Positive Rate: 18.39
- Precision: 72.33

Precision Recall



Threshold value for the model will balance a good precision and recall as 0.41.

Metrics will be created using the threshold 0.41.

Accuracy: 80.58132343846628
Sensitivity: 75.70965125709651
Specificity: 83.5832083958021
False Positive Rate: 16.4167916041979
Precision: 73.96988906497623

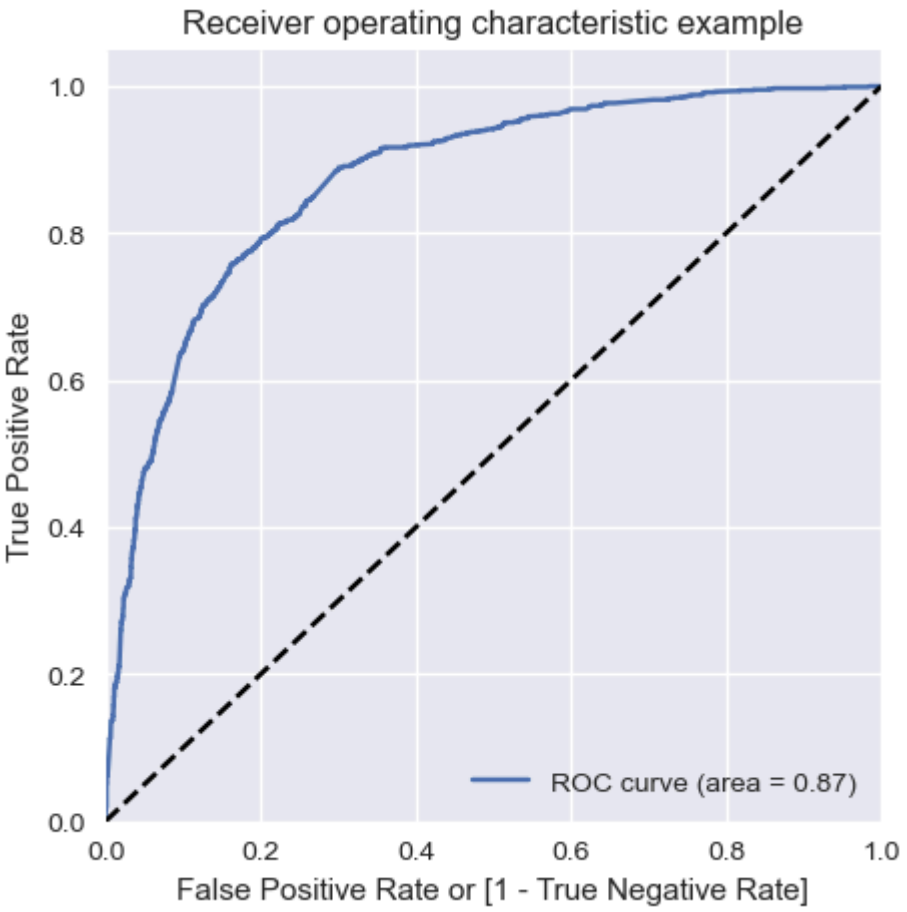
Conclusion: Sensitivity and Specificity is more in the cutoff 0.38, hence the final threshold will be 0.38.

Lead Score Calculation

Lead Score = Converted_Prob * 100.

Prediction on Test data

ROC




```
Accuracy: 80.12265512265512
Sensitivity:  76.71232876712328
Specificity:  82.34943351222421
False Positive Rate:  17.65056648777579
Precision:  73.94366197183099
```

Conclusion:

Metrices on training data and test data set are close.

Hence, the model is performing good.

Summary

- Customers with higher lead scores are more likely to convert and vice versa.
- The sensitivity and accuracy of the model is close to 80 %
- Lead Add Form and SMS Sent are the top features for the consideration of high conversion rate.

Recommendation

- Prioritize features with positive coefficients for targeted marketing.
- Give tailor messages for working professionals, optimize communication channels and aggressively target working professionals for higher conversion rates.