Lead Scoring Case Study

SUBMITTED BY : SHANKAR KOKADWAR AND SHUBHANGI TRIVEDI

Problem Statement and Objective

Business Objective :

- An education company named X Education sells online courses to industry professionals. We need to identify Hot Leads i.e. the leads that are most likely to convert into paying customers.
- To build a logistic regression model to assign a Lead Score value between 0 to 100 to each of the lead which can be used by the company to target potential candidate.

The objective is thus classified into following sub goals:

- Create a logistics regression model to predict the Lead Conversion probabilities for each lead.
- Decide on a probability threshold value above which lead will be predicted as converted
- Multiply the lead conversion probability to arrive at lead score value for each lead.

Problem Solving Methodology

The entire case study has been divided into multiple check- points to achieve all the sub goals . The check points we are using has been listed below:

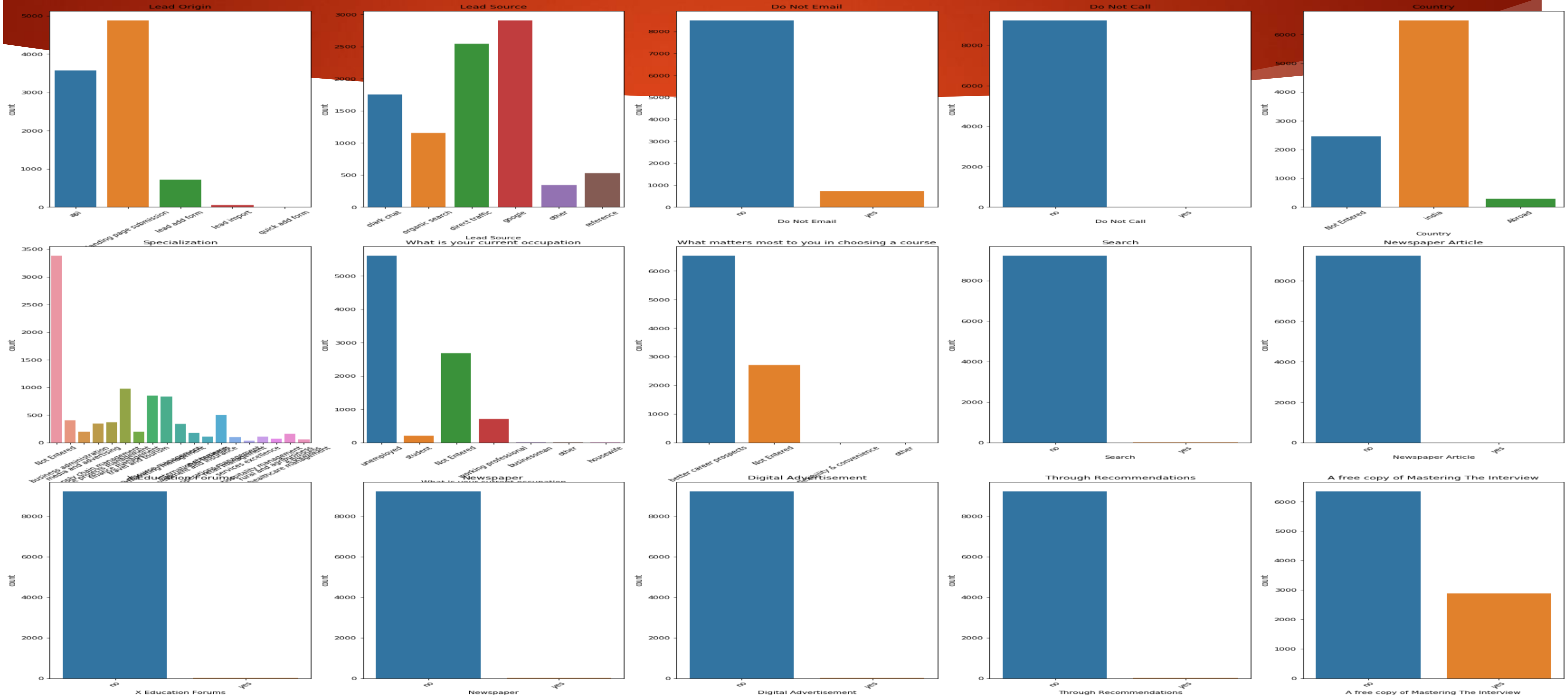
- **Understanding the data set and Data Preparation**.
- **Exploratory Data Analysis** : Univariate Analysis /Bivariate Analysis
- **Dummy Variable Creation**
- **Train – Test Split**
- **Feature Selection :**
  - Applying Recursive Feature elimination to identify 20 best performing subset
  - Building the model with RFE and then manually eliminating features based on high p Value and VIF.
- **Model Building**
- **Model Evaluation:**
  - Perform model evaluation with various metrics like sensitivity, specificity, precision, recall etc.
  - Decide probability based on optimal cut off point and predict the dependent variables for the training data.
  - Use the model for prediction of data set and perform model evaluation on test set.
- **Final conclusion**

Data Cleaning and Preparation

Following data preparation processes were applied to make data dependable, so it provide significant business value:

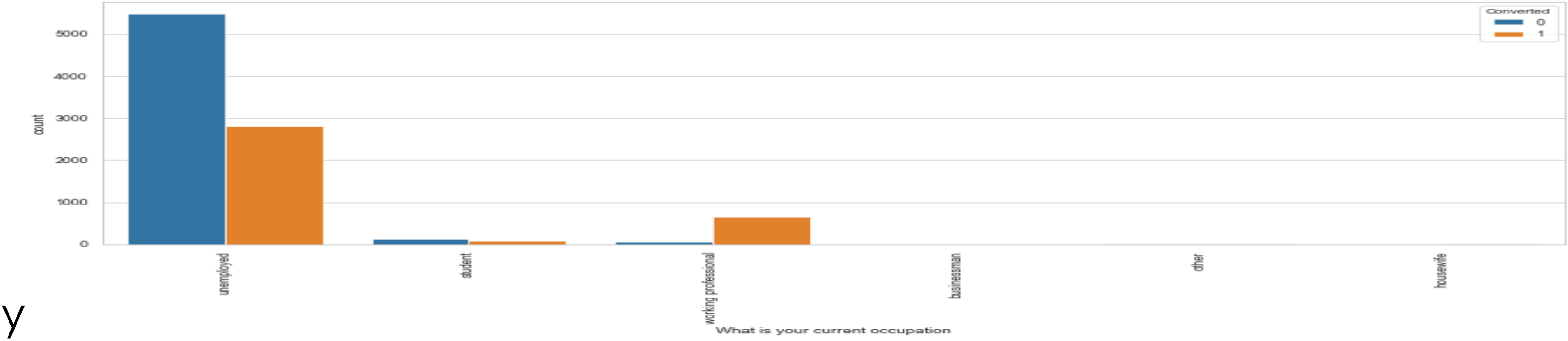
- Removing columns with only 1 unique value
- Some column have **‘Select**’ value and this means that nothing was selected for this column and is equivalent to Null, hence **‘Select’ was replaced by null**.
- There are few columns that have relatively high value of nulls and thus these columns cannot help in prediction, hence columns with more that **39% null values** were **dropped**.
- There were few categorical columns that were **highly skewed**, like 'What matters most to you in choosing a course’ **were dropped.**
- Few categorical columns which have a high number of category but some categories had a relatively less number of rows and **hence we combined categories with less percentage of rows into a single category** e.g.. **'Specialization‘**, **'Lead Source‘**, **‘Last Activity**’
- Columns with **less percentage of missing values were imputed**.
- Columns with **outlier** were subjected to outlier treatment e.g.. ‘TotalVisits’, ‘Page view per visit’

EDA: Univariate Analysis : Through Visualization Categorical Variables

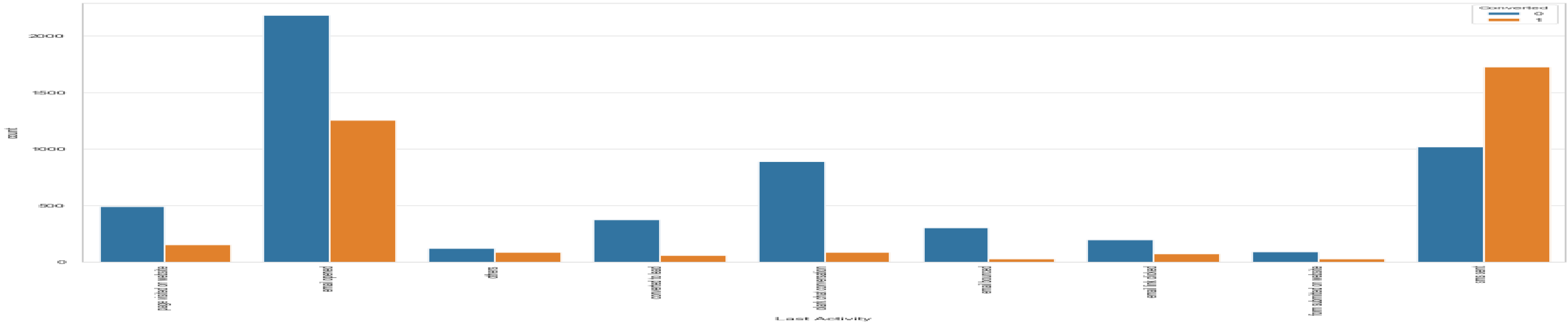


EDA: Bivariate Analysis

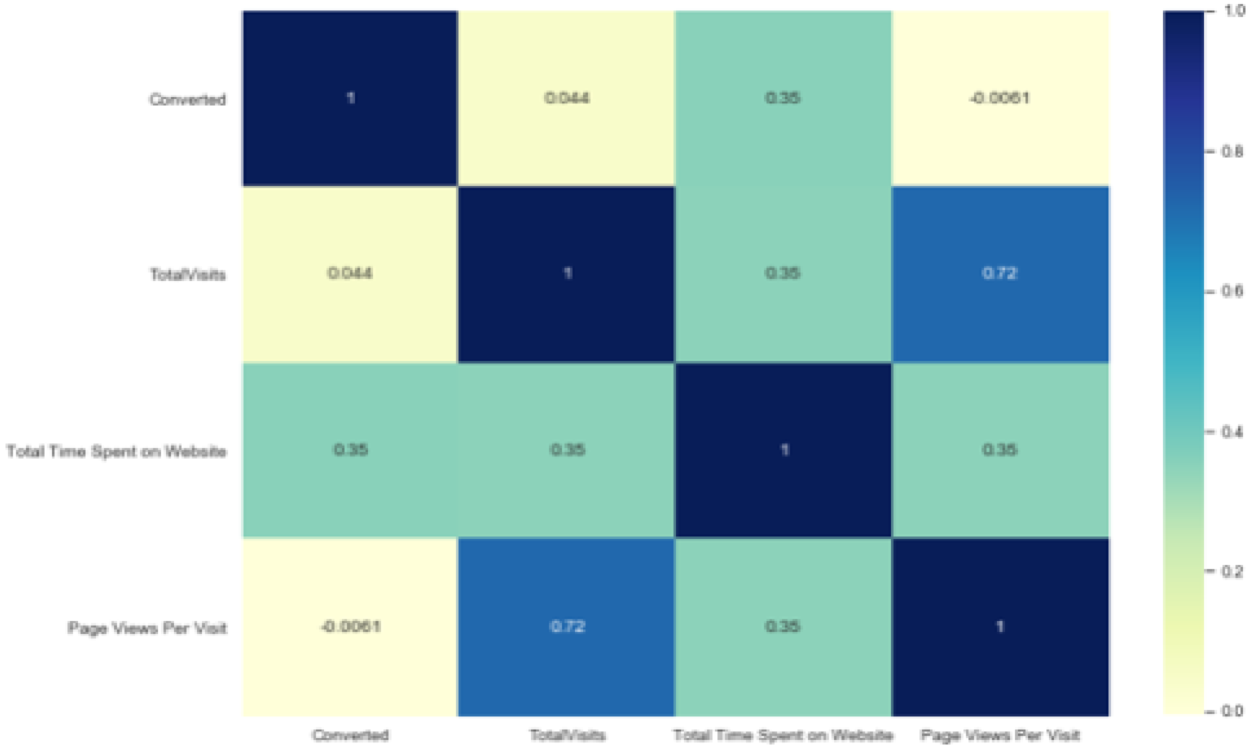
What is your current Occupation



Last activity



EDA: Multivariate Analysis: Correlation Among Numerical Vars



	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Converted	1	0.044	0.35	-0.0061
TotalVisits	0.044	1	0.35	0.72
Total Time Spent on Website	0.35	0.35	1	0.35
Page Views Per Visit	-0.0061	0.72	0.35	1

As we can see that there is not much correlation among the numerical variables

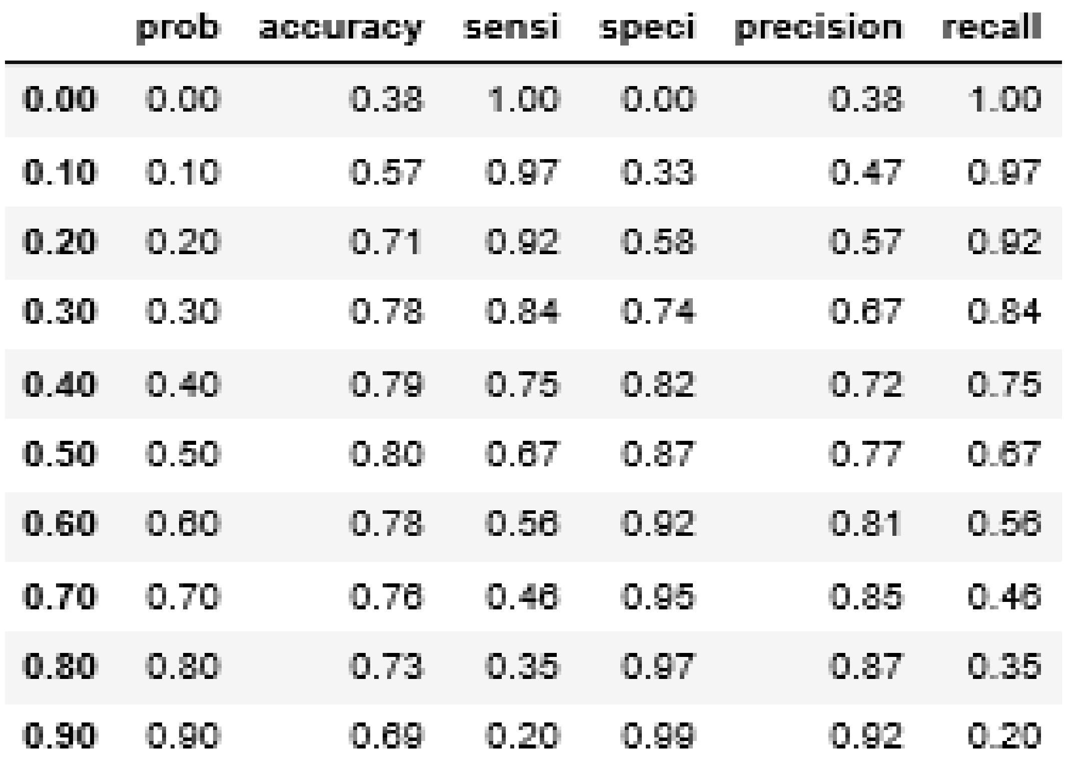
Imbalance for Converted variables

- There are 61.5 percent rows for which converted value is zero
- There are 38.5 percent rows for which converted value is one
- So we can say that data is relatively balanced

ML Algorithm Selection and Approach

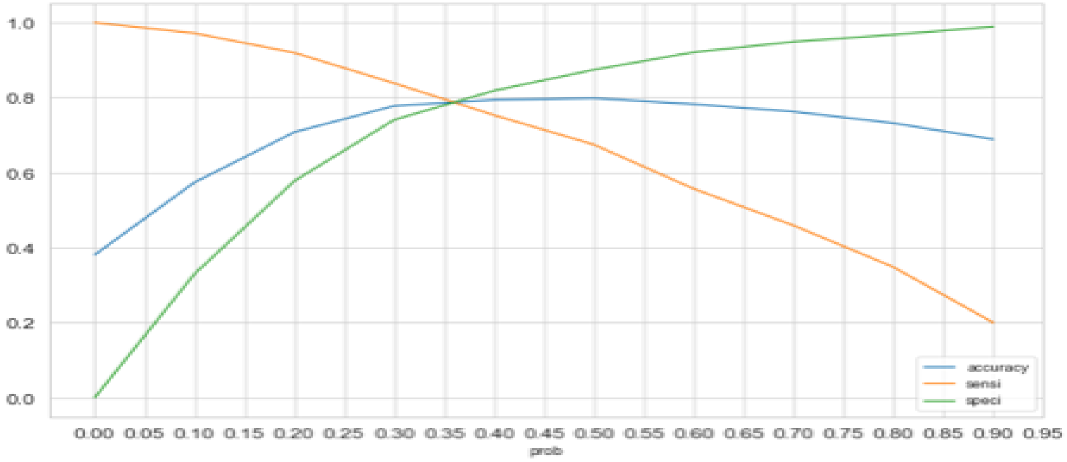
- As we have to create a model where in we have to do binary classification of a lead based on data, logistic regression is suitable for such model.
- **Dummy Variable creation**: The dummy variables were created using **one Hot encoding** for all categorical columns and later original categorical columns were removed.
- **Feature Scaling**: For numeric values we used the **MinMaxScaler** for columns, 'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'
- **Test- Train split of data :** The split was done at 70% and 30% for train and test data respectively.
- **Feature Selection**: we used hybrid approach wherein first we found the 20 most significant variables through RFE and then used manual approach to remove variables one by one based on P value and VIF.
- **Model Building**: By selecting appropriate Probability cutoff value

Model Evaluation: Finding Optimal Probability Threshold Value



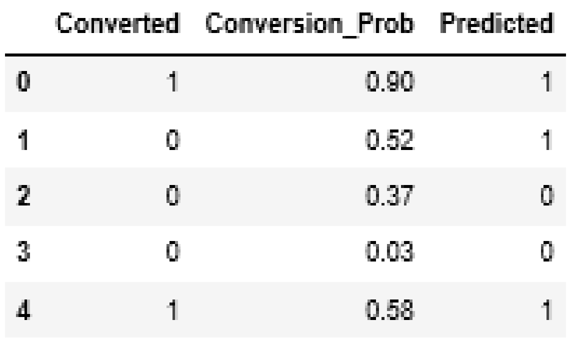
	prob	accuracy	sensi	speci	precision	recall
0.00	0.00	0.38	1.00	0.00	0.38	1.00
0.10	0.10	0.57	0.97	0.33	0.47	0.97
0.20	0.20	0.71	0.92	0.58	0.57	0.92
0.30	0.30	0.78	0.84	0.74	0.67	0.84
0.40	0.40	0.79	0.75	0.82	0.72	0.75
0.50	0.50	0.80	0.67	0.87	0.77	0.67
0.60	0.60	0.78	0.56	0.92	0.81	0.56
0.70	0.70	0.76	0.46	0.95	0.85	0.46
0.80	0.80	0.73	0.35	0.97	0.87	0.35
0.90	0.90	0.69	0.20	0.99	0.92	0.20

- The accuracy, specificity and sensitivity were calculated for various value of probability threshold and plotted in graph.
- From the curve, 0.34 is found to be optimum point for cut-off Probability.
- Sensitivity and specificity is approx. 80%, which is well acceptable value.



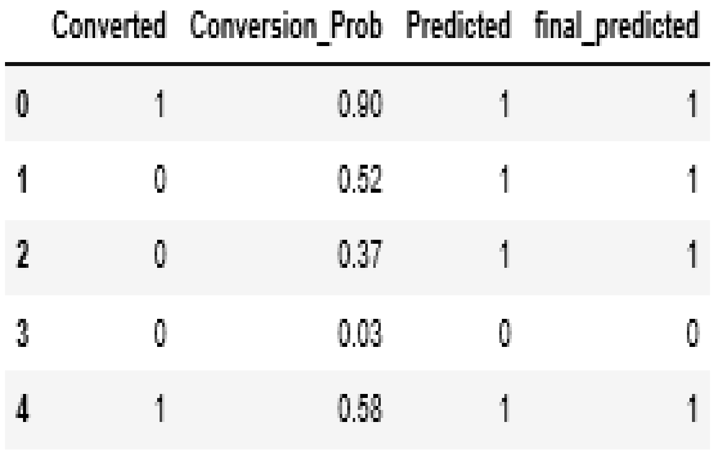
Predicting the conversion probability and Predicted column

Creating a data frame with actual converted flag and predicted probabilities



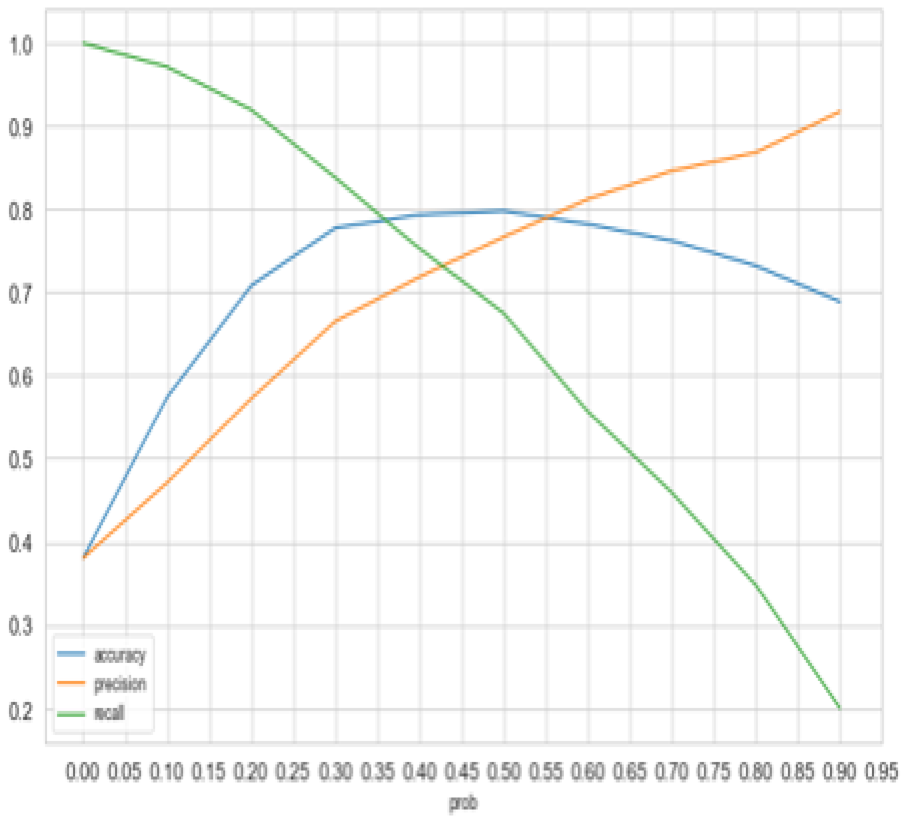
	Converted	Conversion_Prob	Predicted
0	1	0.90	1
1	0	0.52	1
2	0	0.37	0
3	0	0.03	0
4	1	0.58	1

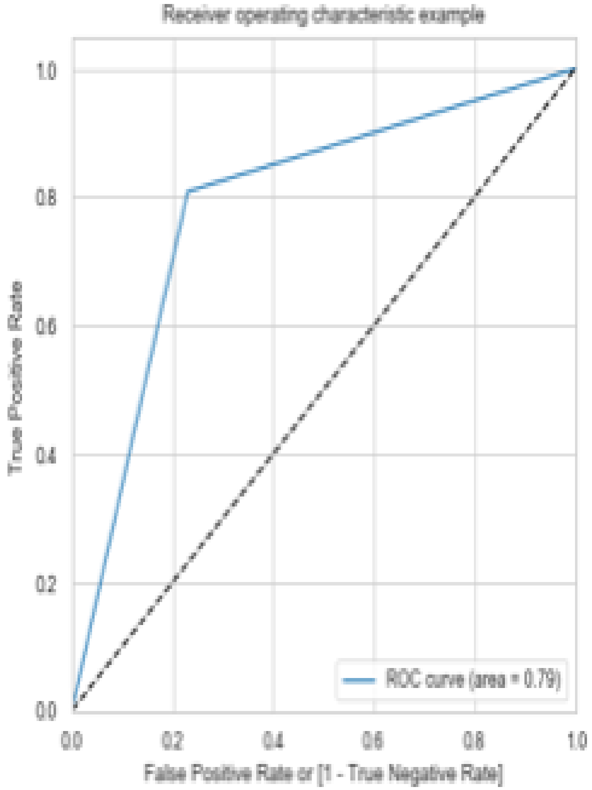
Creating new column final_predicted with 1 if conversion_prob is greater than 0.34 else 0



	Converted	Conversion_Prob	Predicted	final_predicted
0	1	0.90	1	1
1	0	0.52	1	1
2	0	0.37	1	1
3	0	0.03	0	0
4	1	0.58	1	1

Trade offs between Various Model Evaluation Methods and ROC curve





Train Data

Observation: So as we can see above the model seems to be performing well. The ROC curve has a value of 0.80, which is very good. We have the following values for the Train Data:

- Accuracy : 78.45%
- Sensitivity : 80.83%
- Specificity : 76.99%

Test Data

Observation: So as we can see above the model seems to be performing well. The ROC curve has a value of 0.80, which is very good. We have the following values for the Test Data:

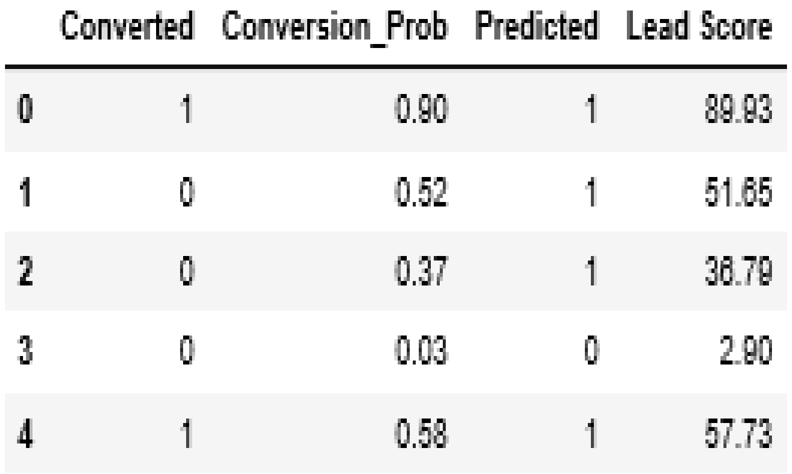
- Accuracy : 79.15%
- Sensitivity : 81.45%
- Specificity : 77.74%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

Lead Score Calculation

Lead score is calculated for all the leads.
Lead score formula is as below :

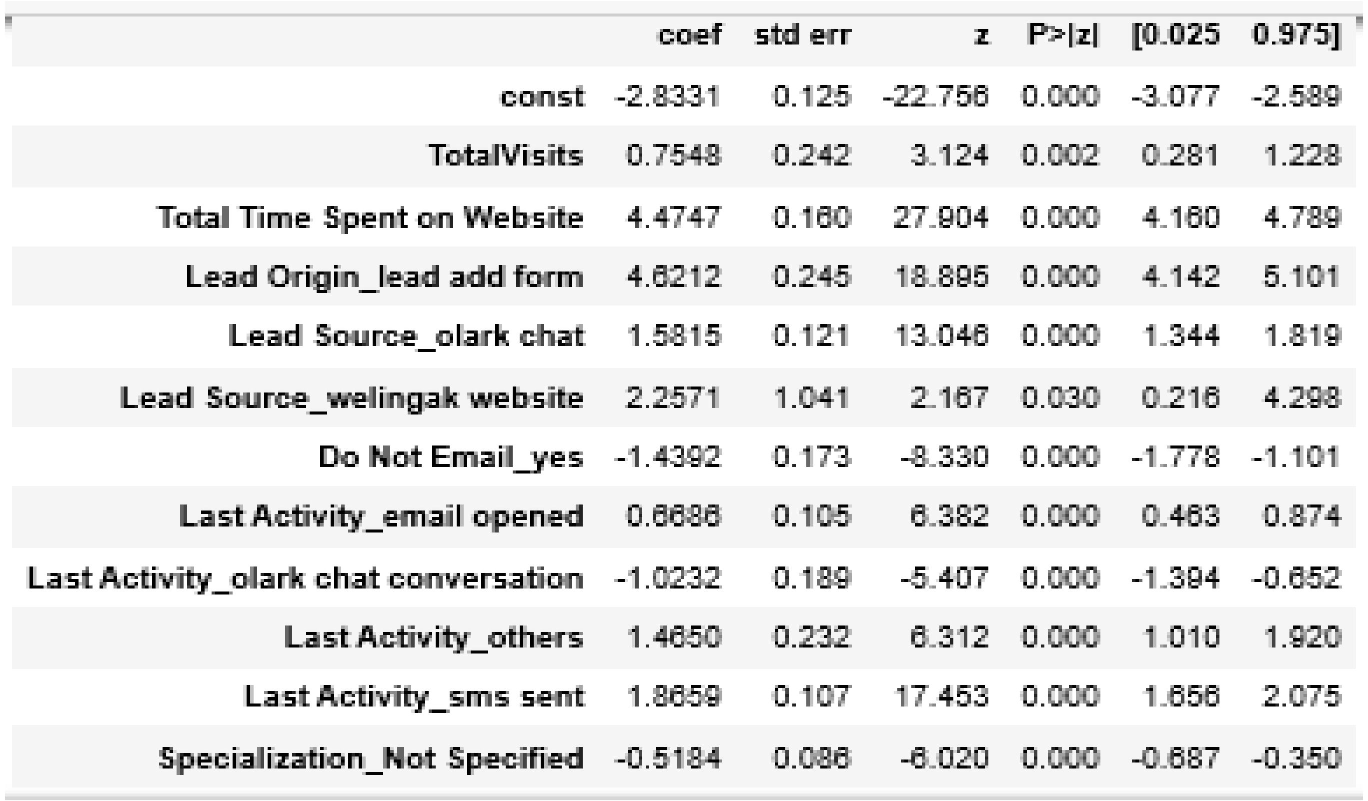
Lead Score = 100 * Conversion Probability



	Converted	Conversion_Prob	Predicted	Lead Score
0	1	0.90	1	89.93
1	0	0.52	1	51.65
2	0	0.37	1	36.79
3	0	0.03	0	2.90
4	1	0.58	1	57.73

- Train and test set is concatenated to get entire list of leads available.
- The conversion rate is multiplied by 100 to get Lead Score
- Higher the lead score, higher is the probability of lead getting converted.
- Since we have selected 0.34 as final probability threshold for deciding if lead is will convert or not , any lead with score above 34 will give 1 for predicted column otherwise it will be 0.
- We have received precision of 68 %, which means for 68% times predicted value will be accurate.

Final Model Summary



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8331	0.125	-22.756	0.000	-3.077	-2.589
TotalVisits	0.7548	0.242	3.124	0.002	0.281	1.228
Total Time Spent on Website	4.4747	0.160	27.904	0.000	4.160	4.789
Lead Origin_lead add form	4.6212	0.245	18.895	0.000	4.142	5.101
Lead Source_olark chat	1.5815	0.121	13.046	0.000	1.344	1.819
Lead Source_welingak website	2.2571	1.041	2.167	0.030	0.216	4.298
Do Not Email_yes	-1.4392	0.173	-8.330	0.000	-1.778	-1.101
Last Activity_email opened	0.6686	0.105	6.382	0.000	0.463	0.874
Last Activity_olark chat conversation	-1.0232	0.189	-5.407	0.000	-1.394	-0.652
Last Activity_others	1.4650	0.232	6.312	0.000	1.010	1.920
Last Activity_sms sent	1.8659	0.107	17.453	0.000	1.656	2.075
Specialization_Not Specified	-0.5184	0.086	-6.020	0.000	-0.687	-0.350

Conclusion and Recommendation:

If Lead Origin is add form then probability of lead getting converted increases.

The students spending more time on the website have the higher chances of being converted.

If Lead source is welingak Website and olark chat, then probability of conversion is high.

The students visiting the platform more no of times have the higher chances of being converted.

If last activity is email opened, that means lead have shown some interest and is more likely to be converted.

If the specialization is not specified then chances of conversion to Lead become very low but from EDA we found out that Working professionals are more likely to join the course so company can create some campaign for working professionals.

Those candidate who are opting out of email service are less likely to be converted.