Lead Scoring Case Study

- worked complied by

Hrishikesh Pradhan

Hitesh Padal

Imtihazahmad Mullanavar

Batch ID: 5705 (DS C67)

Contents

- Problem Statement
  - Business Brief
  - Business Problem
  - Data Brief
  - Goal Brief
- Exploratory Data Analysis (EDA)
  - Procedure
  - Data Handling
  - Insights
- Model Build Procedure
- Final Model
  - Characteristics
  - Metrics
- Test data Predictions & Conclusions
- Business Recommendations

Problem Statement – Business Brief

- 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%.

Problem Statement – Business Problem

- 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.
- A typical lead conversion process can be represented using the funnel (*image*).
- Based on the funnel, 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, it is required to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Lead Conversion Process
- Demonstrated as a funnel

Problem Statement – Data Brief

- We have been provided with a leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
- More about the dataset available from the data dictionary provided in the zip folder at the end of the page.
- Another thing that is required to check out are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (because that means the value has not been selected and a default value was taken as 'Select').

Problem Statement – Goal Brief

Business Goal:

- To help X Education select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Technical Procedure:

- 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 for which, model should be able to adjust to if the company's requirement changes in the future, therefore will need to handle these as well. These problems are provided in a separate doc file and will be answered based on the logistic regression model built in the first step.

Exploratory Data Analysis (EDA) – Procedure

- To identify
  - Volume of data
  - Attributes/Columns in the dataset
    - & Cleaning the attribute data
      - Imputing missing values wherever applicable (Categorical: Mode; Numeric: Median)
      - Ignoring imputation if the proportion of missing values is very high (ex: more than 20% missing data)
      - Text based attributes: Updating the values to proper case / correcting spellings as applicable to give meaning to values and avoid duplicates in the unique values list.
- To analyse the attributes
  - To understand the distribution of data in the attributes
  - To understand the distribution proportional of data in the attributes w.r.t. the 'Converted' attribute(s)
    - Identify the abnormal distribution among the attribute values, and stating the reasons if insightful
    - Identify the proportion of distribution w.r.t. 'Target' and flagging/stating such attribute values as segments of attention and further actions to be taken by business team as applicable to reduce/avoid unfavourable criteria

EDA – Data Handling (Leads.csv)

- All the observations in this process (data handling) and actions taken are stated adjacent to the step/cell where such action is performed.
- Some imputations may be based on assumption/data understanding, rather than empirical method(s).
- Step by step actions taken:
  - Removed 'Prospect ID' and 'Lead Number'
  - Replaced dummy 'Select' values with nan
  - Obtained count and proportion of null values in each column, large distribution of null values shown below ie., > 40%

How did you hear about X Education	78.46	Asymmetrique Activity Score	45.65
Lead Profile	74.19	Asymmetrique Profile Score	45.65
Lead Quality	51.59	Asymmetrique Profile Index	45.65
		Asymmetrique Activity Index	45.65

  - Removed these attributes to avoid bias in predictions
  - Replaced missing values on the column

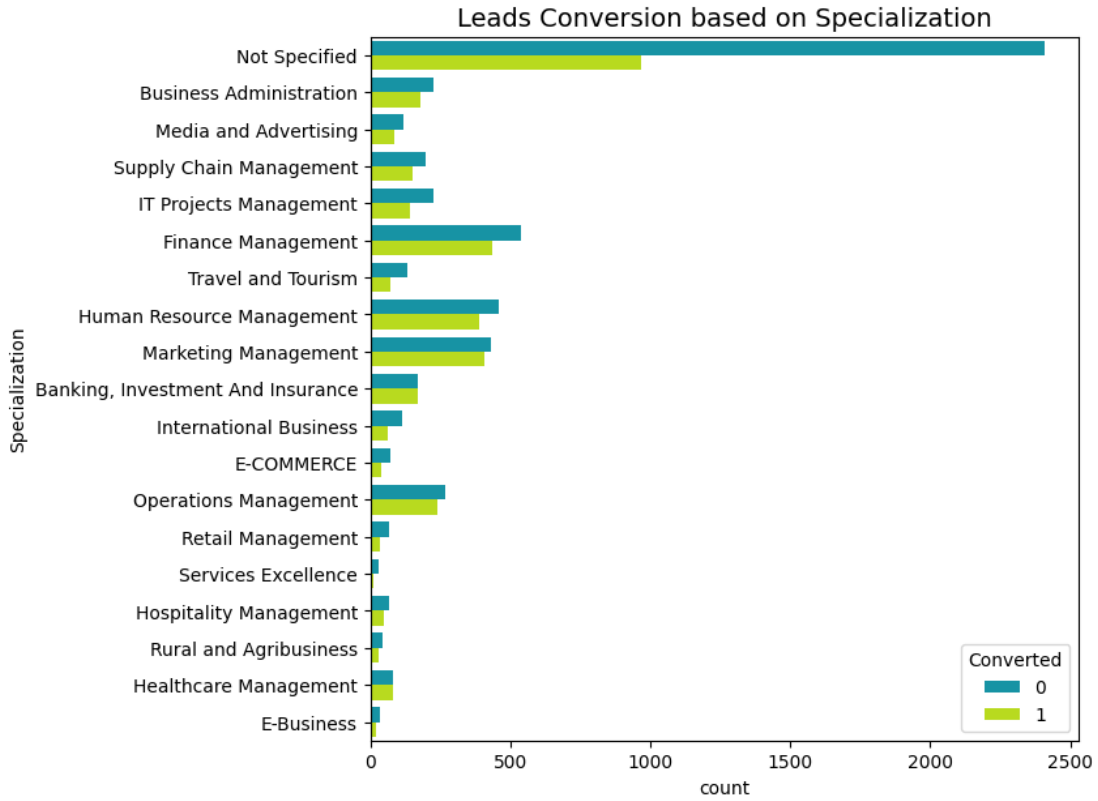
'City' with Mode = 'Mumbai'.	'What is your current occupation' with 'Unemployed'
'Specialization' with 'Not Specified'	'Lead Source' with 'Others'
'Tags' with 'Not Specified	'Last Activity' with 'Others'

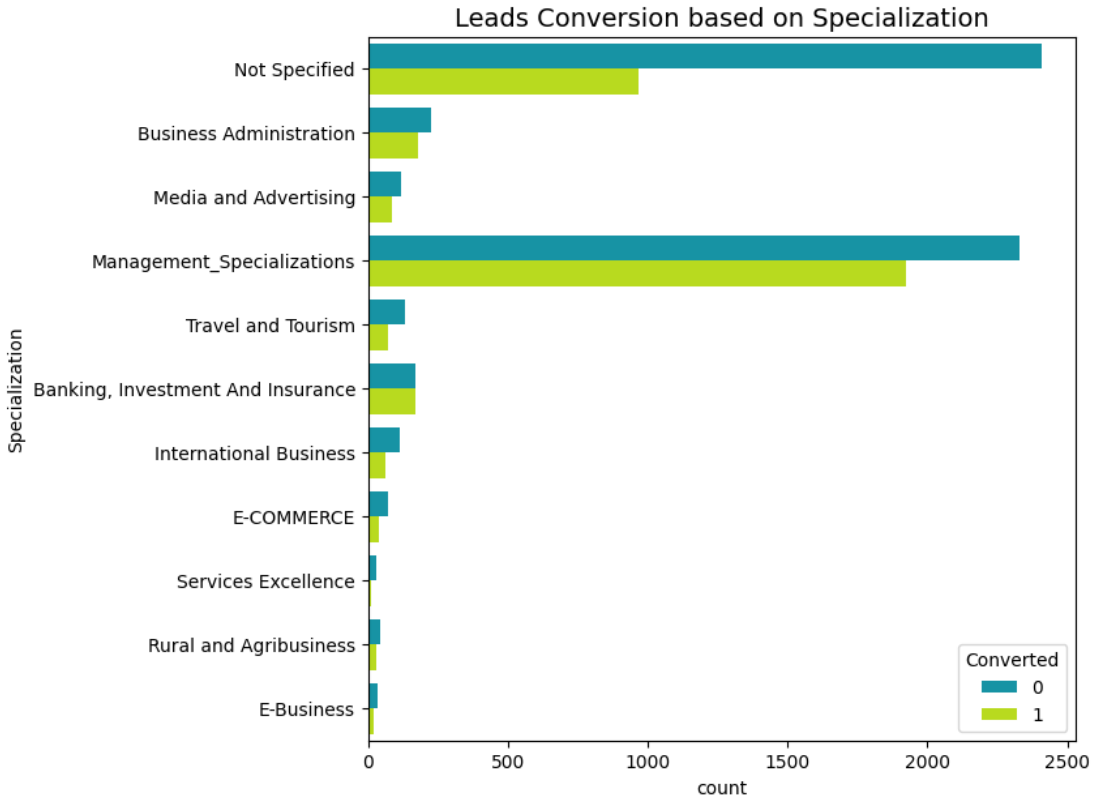
  - Replaced certain low frequent values on the column

'Tags' with 'Other_Tags'	'Lead Source' with 'Others'	'Last Activity' with 'Others'

EDA – Data Insights

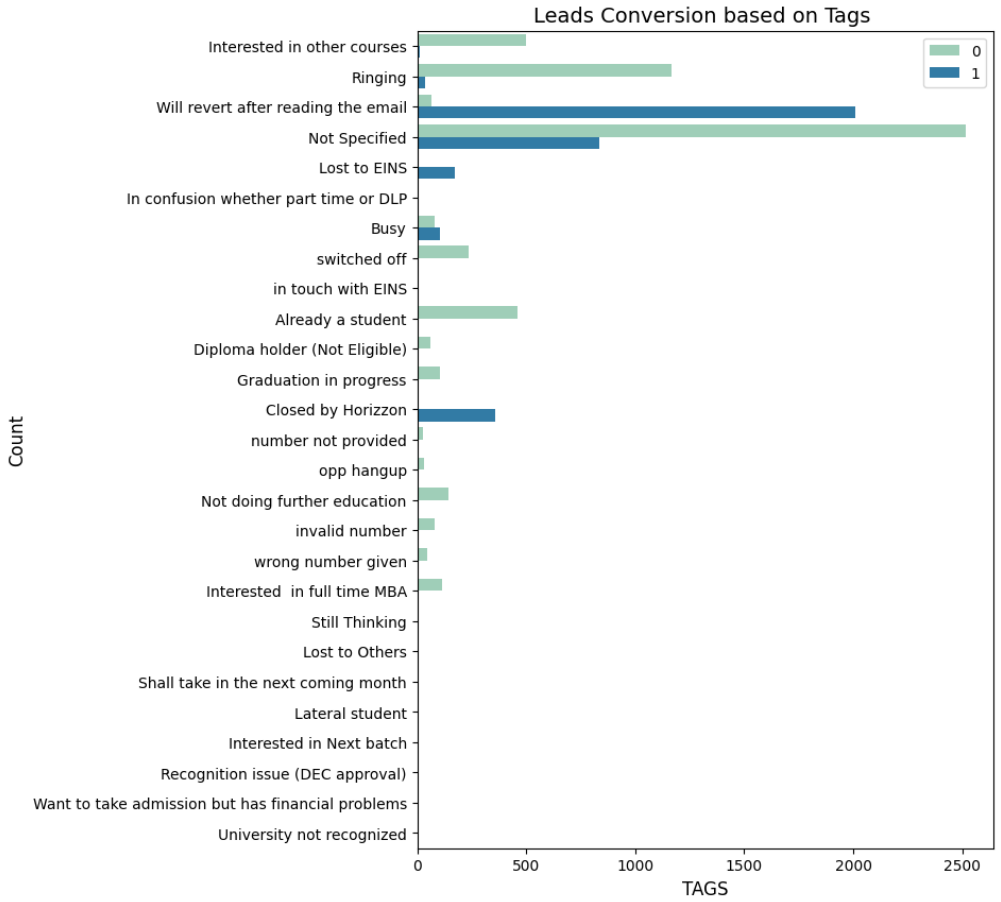
- Based on **Specialization**, more number and significant proportion of leads are converted among these choices:: Management: Finance, Human Resource, Marketing, and Operations.
- After grouping all Management Specializations as one, it has highest number of leads converted.

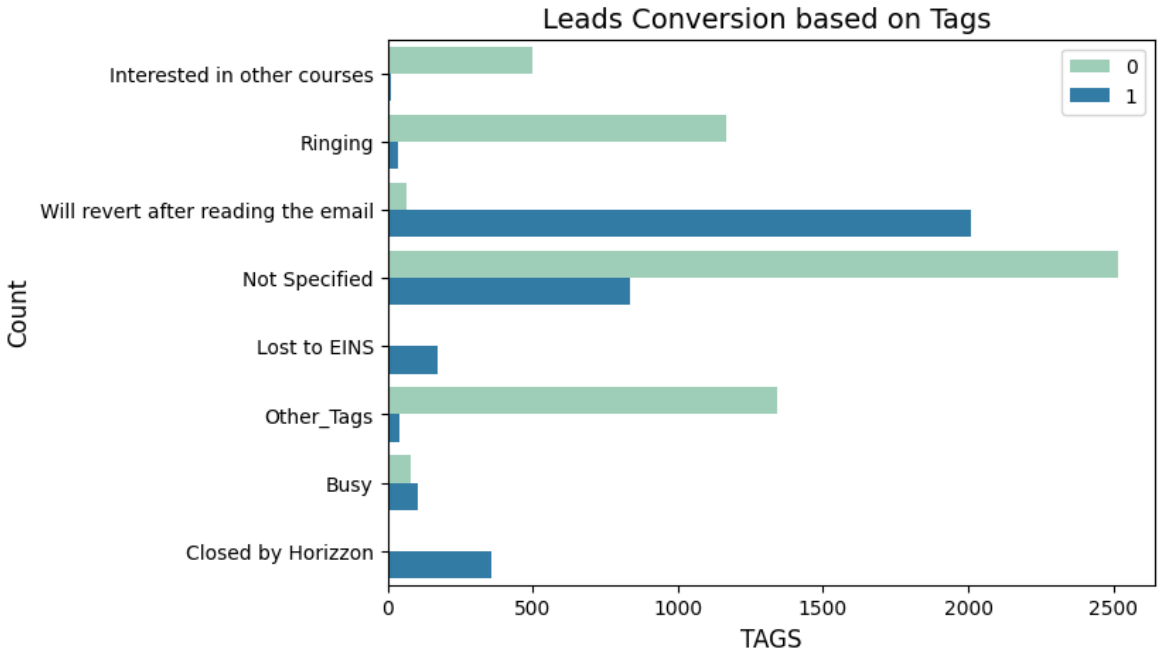




EDA – Data Insights (contd.)

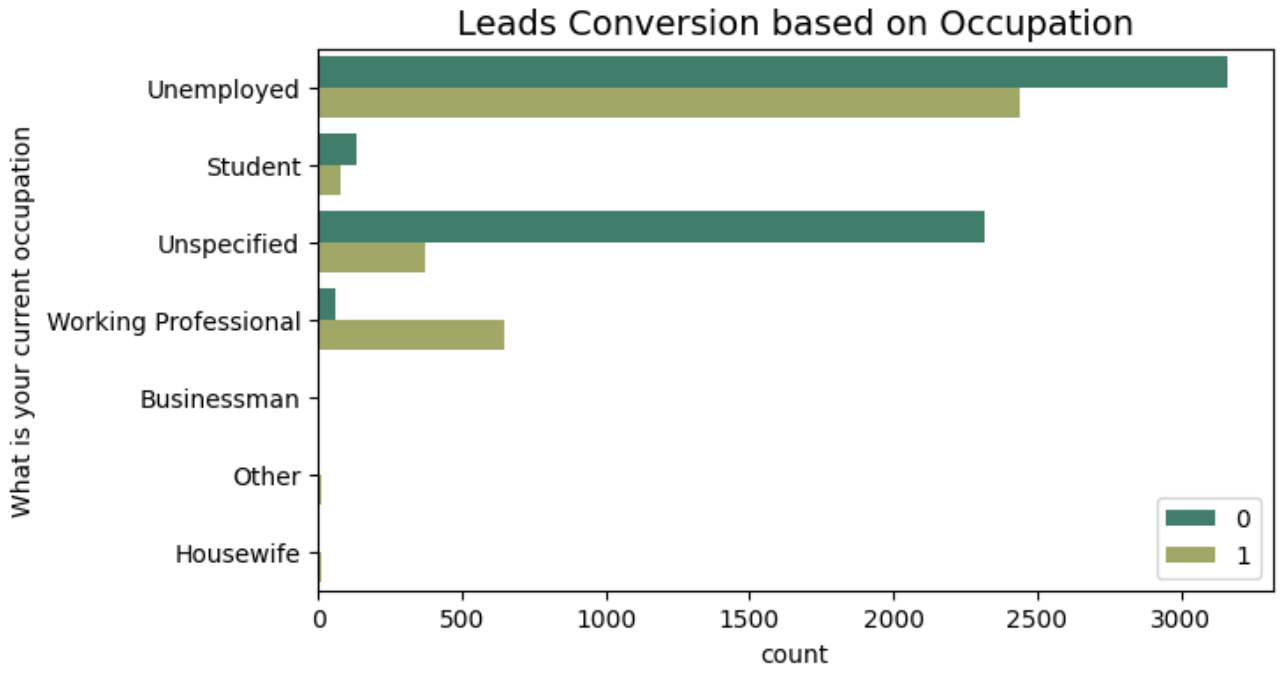
- Based on **Tags**, more number and significant proportion of leads are converted among these choices:: Will revert after reading the email.
- After grouping the very less count categories as one: Other_Tags, we have negligible conversions from this group.





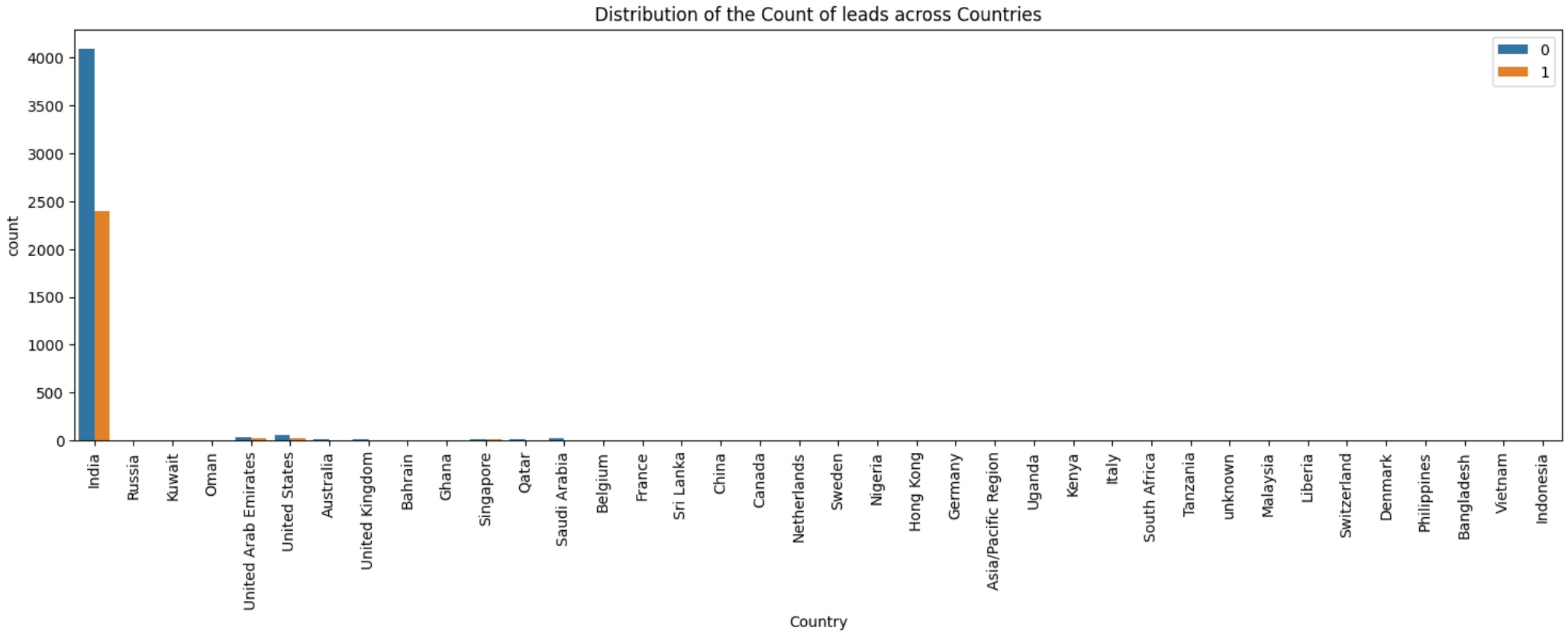
EDA – Data Insights (contd.)

- Based on **Occupation**, more number and significant proportion of leads are converted among these choices:: Unemployed and Unspecified.
- There is a strong likelihood that working professionals will opt for the course.
- The largest group among the leads consists of unemployed individuals, and those who didnot specify any choice.
- Categories like housewives, businessmen, students, and others are less likely to convert and enroll in the course.



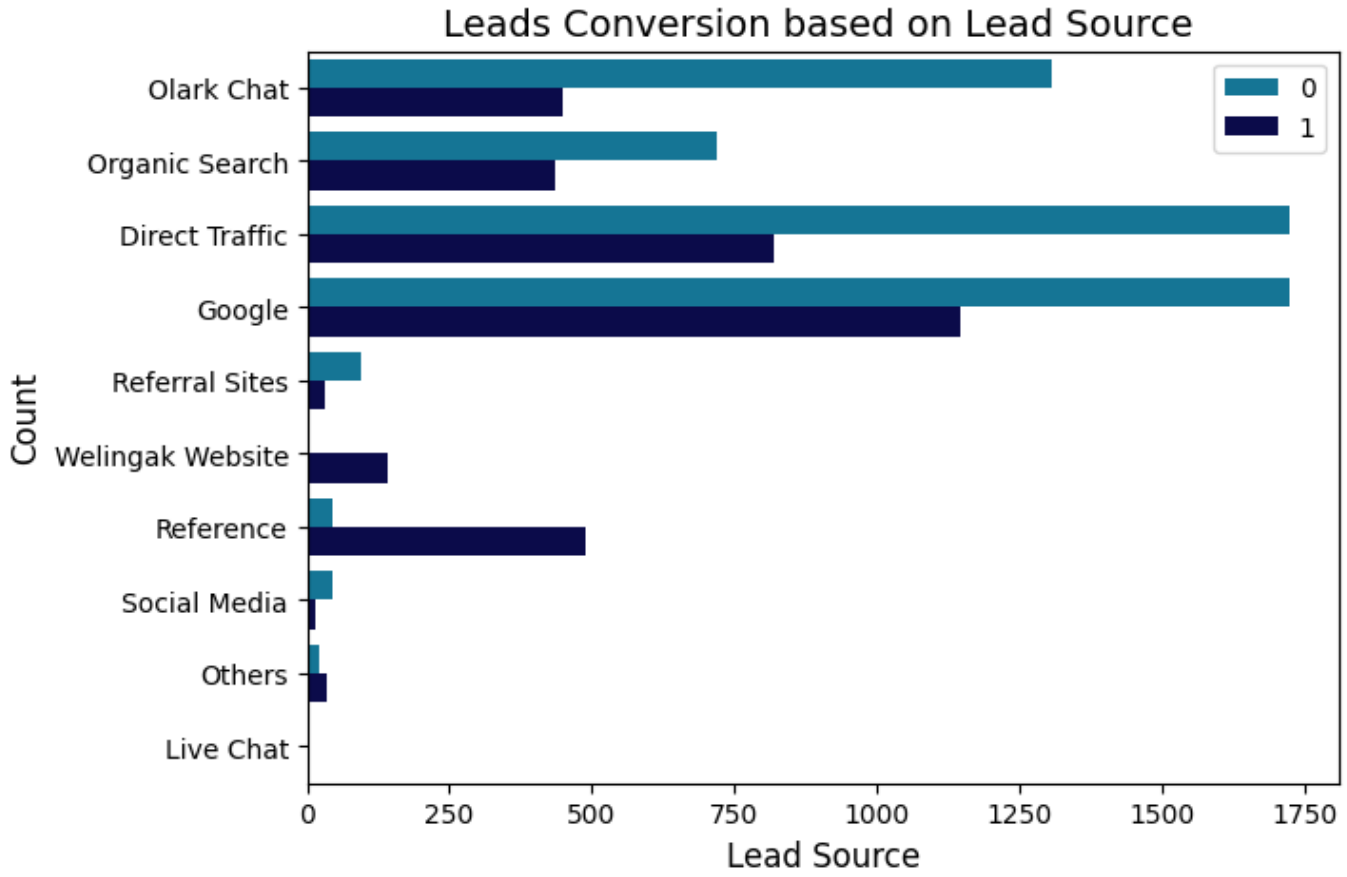
EDA – Data Insights (contd.)

- Based on **Country**, more number and significant proportion of leads are converted among these choices:: India.
- Because "India" shows as the most occurring Country, it may not be suitable for an analysis - especially for a classification problem. Therefore we remove the Country column to avoid the bias and high VIF.



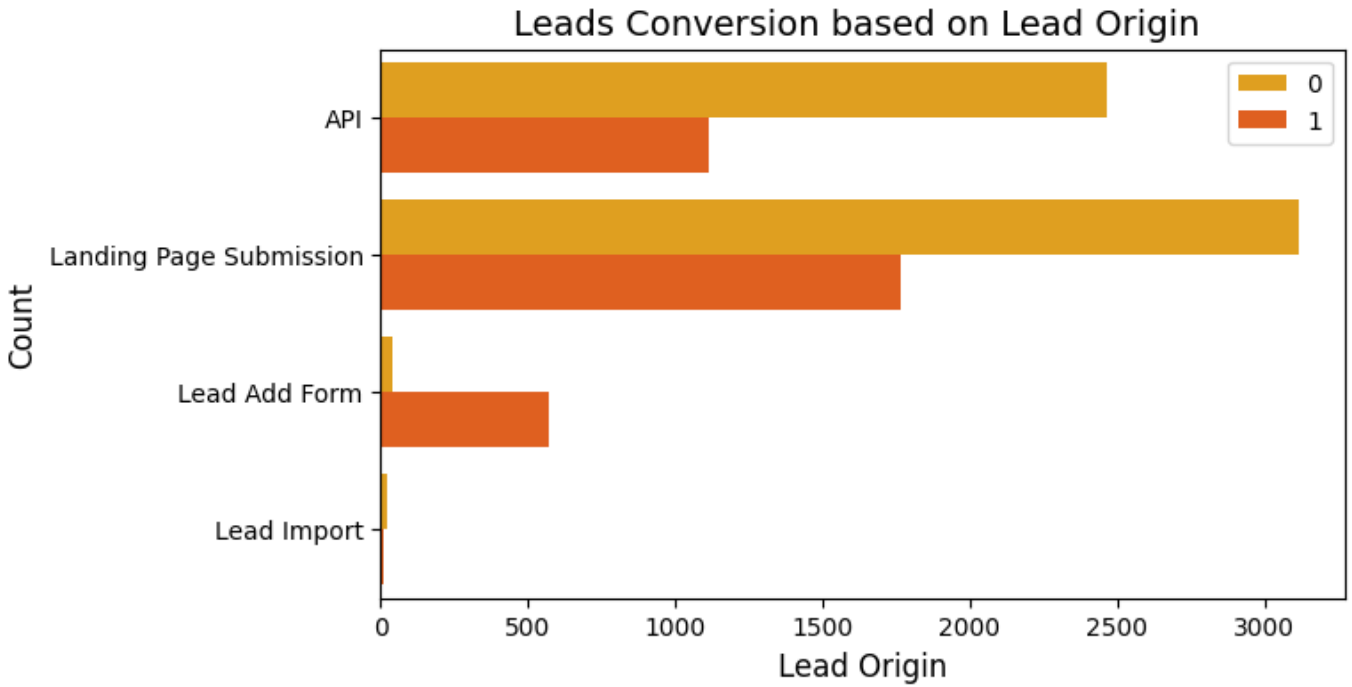
EDA – Data Insights (contd.)

- Based on **Lead Source**,
- The majority of leads are generated through Google and direct traffic, with the fewest coming from live chat.
- The Welingak website has the highest conversion rate.
- Improving lead conversion can be achieved by maximizing leads from references and the Welingak website.
- Focusing on Olark chat, organic search, direct traffic, and Google leads could further boost lead conversion rates.



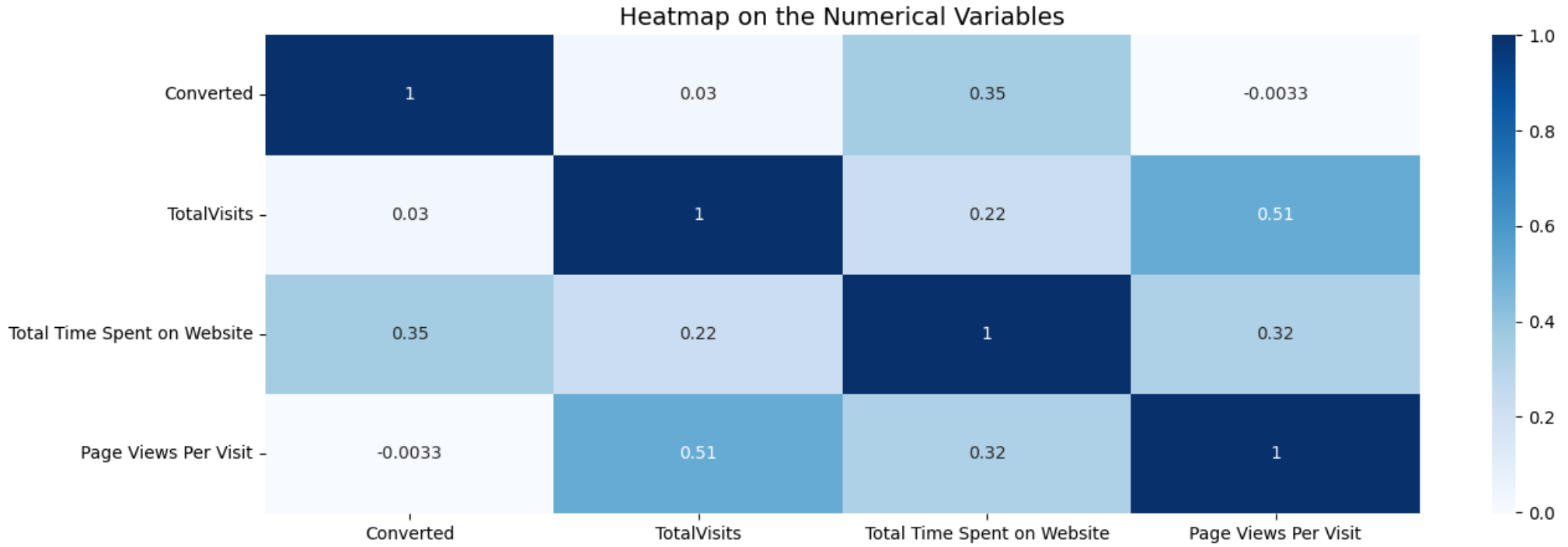
EDA – Data Insights (contd.)

- Based on **Lead Origin**,
- Both API and landing page submissions generate a high volume of leads and conversions.
- While the lead add form has a strong conversion rate, the number of leads it generates is relatively low.
- Increasing the number of leads through the lead add form could significantly boost the overall conversion rate and contribute to greater growth.



EDA – Data Insights (contd.)

- Correlation analysis of numeric variables
- Users who spent more time on website are more likely to get converted with 35% probability.
- Total visits to the website is highly correlated with Total time spent on website, and page views per visit is highly correlated with total visits.



Model Build Procedure

Step by step procedure:

- Convert all object/text based variables to binary variables using dummy variable encoding
- Reduce dimensionality of data using RFE (Recursive Feature Elimination) method and obtain top 20 variables.
- Now train the logistic regression model and check for p-values of the variables and calculate the VIF (Variance Inflation Factor) of these variables.
- Exclude one variable at a time from the model build process, that have p-value more than 0.05 and/or VIF greater than 5, and rerun the above step and feature elimination process until this can not be further done.
- Now perform model evaluation using various metrics such as sensitivity (recall score), specificity, precision score, positive predictive value, negative predictive value.
- Finally, plot and analyse the ROC curve and proceed for obtaining optimal cut-off point for prediction probability. Using this cut-off we can obtain the right conversion prediction from the predicted conversion probability.
- We can now proceed to run the final model against the test data and run metrics to compare and conclude the model performance w.r.t. the training process and provide additional prescriptive points for improving either model building procedure or about data quality or quantity that can be improve model performance.

Final Model – Characteristics

- Final model characteristics
- P-value and VIF are within the assumed cut-offs

  i.e., 0.5 and 5.0 respectively

Features	VIF
Lead Origin_Landing Page Submission	4.32
Page Views Per Visit	3.82
Specialization_Not Specified	2.99
Last Activity_Email Opened	2.65
Last Activity_SMS Sent	2.43
Total Time Spent on Website	2.14
Lead Source_Olark Chat	2.06
TotalVisits	1.97
Last Notable Activity_Modified	1.95
Lead Origin_Lead Add Form	1.70
What is your current occupation_Not Specified	1.61
Lead Source_Welingak Website	1.36
Do Not Email_Yes	1.20
What is your current occupation_Working Profes...	1.20
Last Activity_Others	1.16
Last Notable Activity_Olark Chat Conversation	1.16
Last Notable Activity_Had a Phone Conversation	1.07

Generalized Linear Model Regression Results

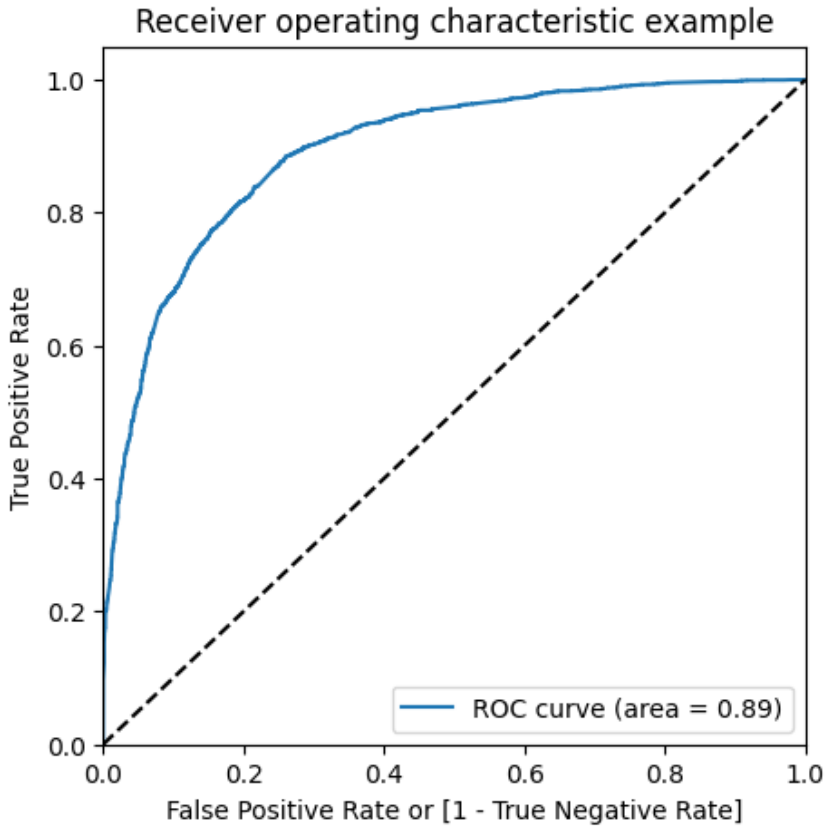
Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6354
Model Family:	Binomial	**Df Model:**	17
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2544.0
Date:	Tue, 22 Oct 2024	**Deviance:**	5088.1
Time:	12:14:44	**Pearson chi2:**	6.29e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4110
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5937	0.177	-8.994	0.000	-1.941	-1.246
TotalVisits	10.8103	2.593	4.169	0.000	5.728	15.893
Total Time Spent on Website	4.5404	0.170	26.696	0.000	4.207	4.874
Page Views Per Visit	-4.0964	1.334	-3.071	0.002	-6.711	-1.482
Lead Origin_Landing Page Submission	-0.8708	0.127	-6.866	0.000	-1.119	-0.622
Lead Origin_Lead Add Form	3.1964	0.243	13.147	0.000	2.720	3.673
Lead Source_Olark Chat	1.0022	0.134	7.505	0.000	0.740	1.264
Lead Source_Welingak Website	2.2216	0.757	2.935	0.003	0.738	3.705
Do Not Email_Yes	-1.3050	0.177	-7.365	0.000	-1.652	-0.958
Last Activity_Email Opened	0.6264	0.113	5.543	0.000	0.405	0.848
Last Activity_Others	1.3583	0.239	5.680	0.000	0.890	1.827
Last Activity_SMS Sent	1.7901	0.115	15.614	0.000	1.565	2.015
Specialization_Not Specified	-0.8367	0.123	-6.819	0.000	-1.077	-0.596
What is your current occupation_Not Specified	-1.1229	0.089	-12.571	0.000	-1.298	-0.948
What is your current occupation_Working Professional	2.4658	0.193	12.805	0.000	2.088	2.843
Last Notable Activity_Had a Phone Conversation	2.0938	1.209	1.732	0.083	-0.276	4.464
Last Notable Activity_Modified	-0.6571	0.091	-7.233	0.000	-0.835	-0.479
Last Notable Activity_Olark Chat Conversation	-0.6921	0.337	-2.051	0.040	-1.354	-0.031

Final Model – Metrics

Prediction metrics based on assumed probability cut-off as 0.5 →

ROC curve based on true conversion probability predicted by the final model




```
# Confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.predicted )
print(confusion)
```

```
[[3498  455]
 [ 710 1709]]
```

```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted))
```

```
0.8171688637790333
```

Metrics beyond simply accuracy

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.7064902852418354
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8848975461674677
```

```
# Calculate false postive rate - predicting churn when customer does not have churned
print(FP/ float(TN+FP))
```

```
0.11510245383253226
```

```
# positive predictive value
print (TP / float(TP+FP))
```

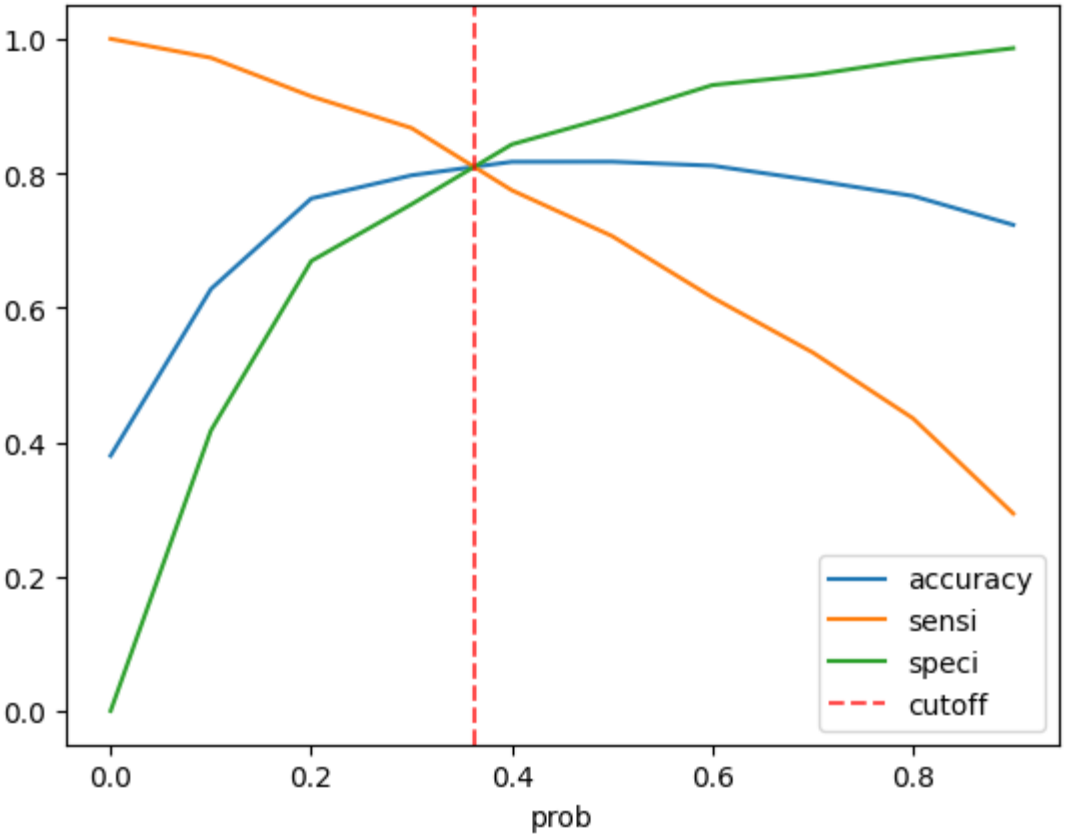
```
0.7897412199630314
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

```
0.8312737642585551
```

Final Model – Metrics

Optimal cut-off for the final model : 0.362



Recalculated performance metrics based on optimal cut-off

```
# Confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_predicted )
print(confusion)
```

```
[[3232  721]
 [ 477 1942]]
```

```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted))
```

```
0.8119899560577527
```

Metrics beyond simply accuracy

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity aka. Recall score of our logistic regression model
TP / float(TP+FN)
```

```
0.8028110789582472
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.8176068808499873
```

```
# Calculate false postive rate - predicting churn when customer does not have churned
print(FP/ float(TN+FP))
```

```
0.18239311915001266
```

```
# positive predictive value aka. Precision score
print (TP / float(TP+FP))
```

```
0.7292527224934284
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

```
0.8713939067133999
```

Test data – Prediction metrics

Test data prediction metrics

```
# Confusion matrix
confusion = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_predicted )
print(confusion)
```

```
[[1644   45]
 [ 651  391]]
```

```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted))
```

```
0.7451482973269864
```

Metrics beyond simply accuracy

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
# Let's see the sensitivity aka. Recall score of our logistic regression model
TP / float(TP+FN)
```

```
0.3752399232245681
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.9733570159857904
```

```
# Calculate false postive rate - predicting churn when customer does not have churned
print(FP/ float(TN+FP))
```

```
0.02664298401420959
```

```
# positive predictive value aka. Precision score
print (TP / float(TP+FP))
```

```
0.8967889908256881
```

```
# Negative predictive value
print (TN / float(TN+ FN))
```

```
0.7163398692810458
```

```
precision_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.8967889908256881
```

```
recall_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.3752399232245681
```

Test data – Conclusions

Final Conclusions and score interpretation

- Based on the model performance metrics, although recall score is fairly low, other metrics are in the desirable range. And we could say that the model predictions are good.
- However, if we want to further improve the model performance in terms of recall as well, we can calculate the above metrics at every iteration of feature elimination (after selecting the feature limit, here considered as 20, at final iteration got 16 features), and stop where further feature elimination is causing significant impact in the model evaluation metrics.

Business Recommendations

- Based on the final model and absolute value of coefficients of the parameters/variables, we can conclude that below are the top 3 variables/features from the data that is ready for modeling.
  - TotalVisits with coefficient of 10.758387
  - Total Time Spent on Website with coefficient of 4.539055
  - Page Views Per Visit with coefficient of 4.070615 (absolute), if considering non-absolute value then based on +ve sign, top 3rd variable is Lead Origin_Lead Add Form with coefficient of 3.197021
- Based on RFE procedure (provided at the end of python (.ipynb file)), we obtained that below are the top 3 variables features from the data that is ready for modeling.
  - Total Time Spent on Website
  - Lead Origin_Lead Add Form
  - What is your current occupation_Working Professional
- The effective strategy would be to choose the leads based on the coefficient/weight of the variable, ie., giving higher priority to those leads who are positive on the highest weighted variable then gradually moving towards the next top weighted vairables in the order of descending order of weight of variables.

Thank you