LEAD SCORING CASE STUDY

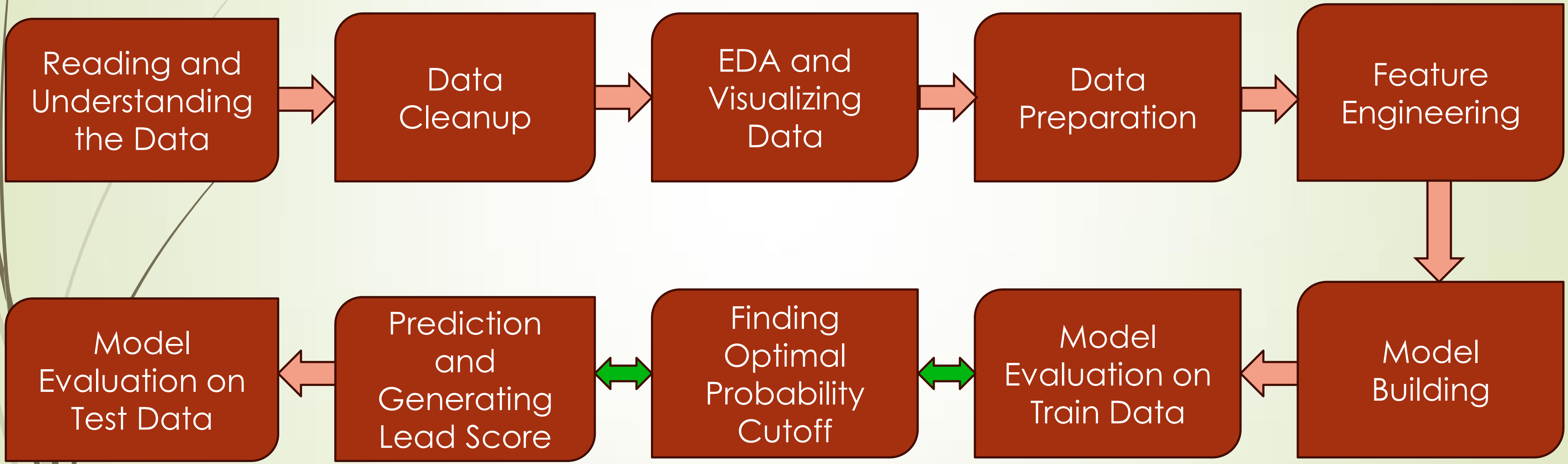
BUSINESS PROBLEM STATEMENT

- 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%.

GOAL

- To identify the features that contributes to predict Lead Conversion.
- Identifying Hot Leads by generating Lead Score for all leads, so that leads having higher Lead Scores can be contacted with priority for achieving Higher Lead Conversion Rate.

OVERALL APPROACH



Reading and Understanding the Data

Data Cleanup

EDA and Visualizing Data

Data Preparation

Feature Engineering

Model Building

Model Evaluation on Train Data

Finding Optimal Probability Cutoff

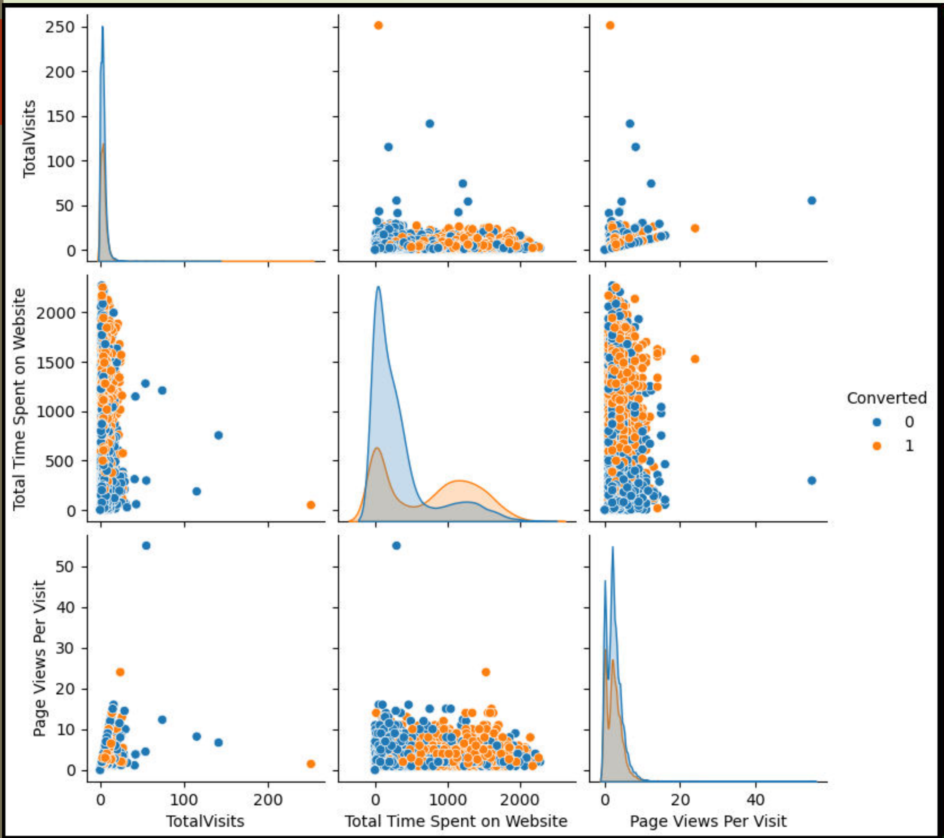
Prediction and Generating Lead Score

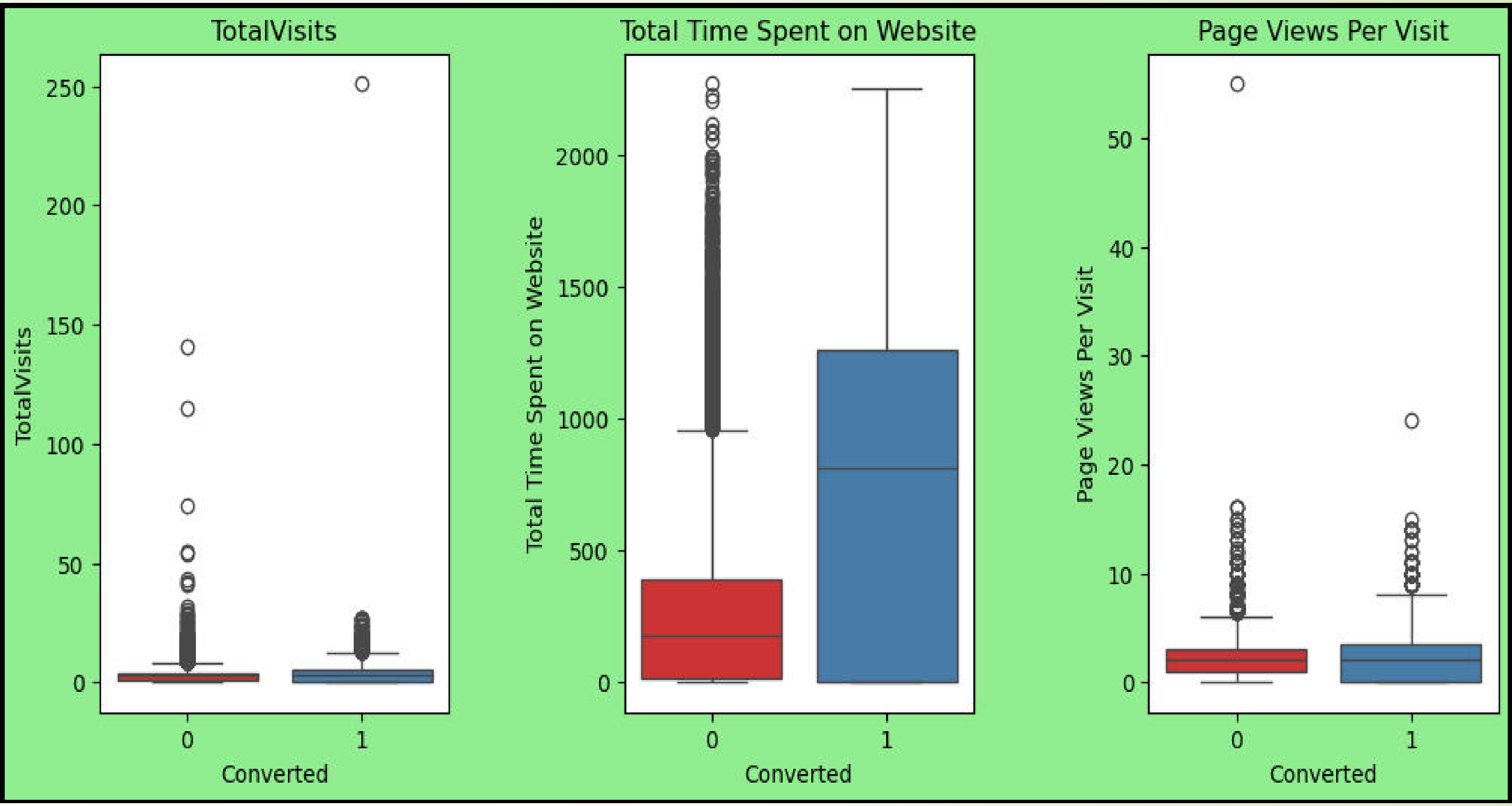
Model Evaluation on Test Data

Reading and Understanding Data & Data Cleanup

- The dataset consists of 37 features in total, where 30 of them are categorical variables and 7 are numerical. It contains 9240 records or data points.
- The class "Select" appears in various columns such as Specialization, How did you hear about X Education, Lead Profile, and City.
- Since "Select" is not a valid class, it likely represents the default value in the form dropdown. If a user didn't choose an option, the value stayed as "Select." We replaced these instances with NaN.
- Lead Quality, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score - These columns have more than 40% missing value. So, we have dropped these columns.
- Missing value treatment was done based on business understanding. For the columns Specialization, What is your current occupation, 'What matters most to you in choosing a course, Tags, City , NaN values were replaced with a new category called "Others."

Visualizing Data & EDA : Numerical variables

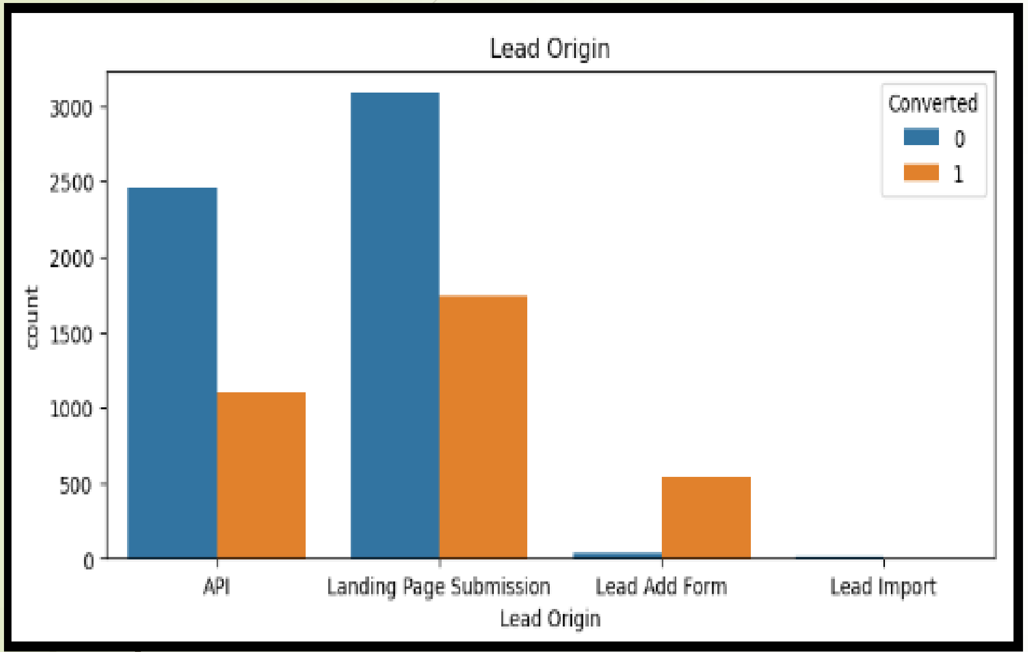


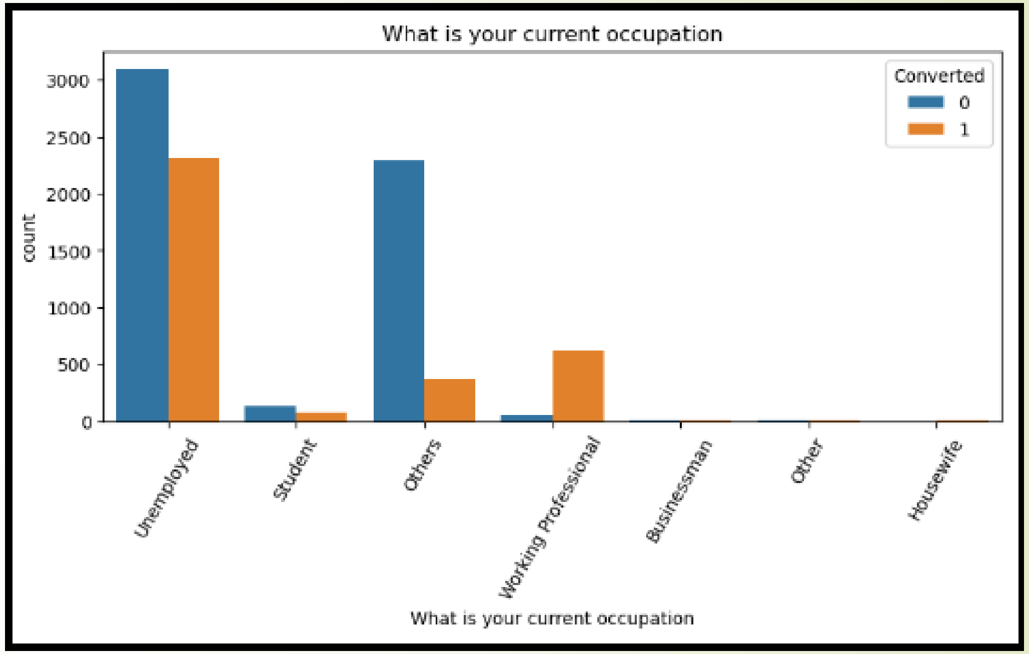


INSIGHTS:

- The median value of ‘Total Time Spent on Website’ is significantly higher for converted leads compared to non-converted ones. The team should focus on targeting customers who spend more time on the website, as they have a greater likelihood of conversion.
- There are many outliers in 'TotalVisits' for leads that were not converted. A significant number of customers are visiting the website frequently but aren't enrolling in the course. The team should investigate why this is happening. Possible reasons could include financial constraints, searching for courses not currently offered by X Education, or finding better alternatives from competitors.

Visualizing Data & EDA : Categorical variables

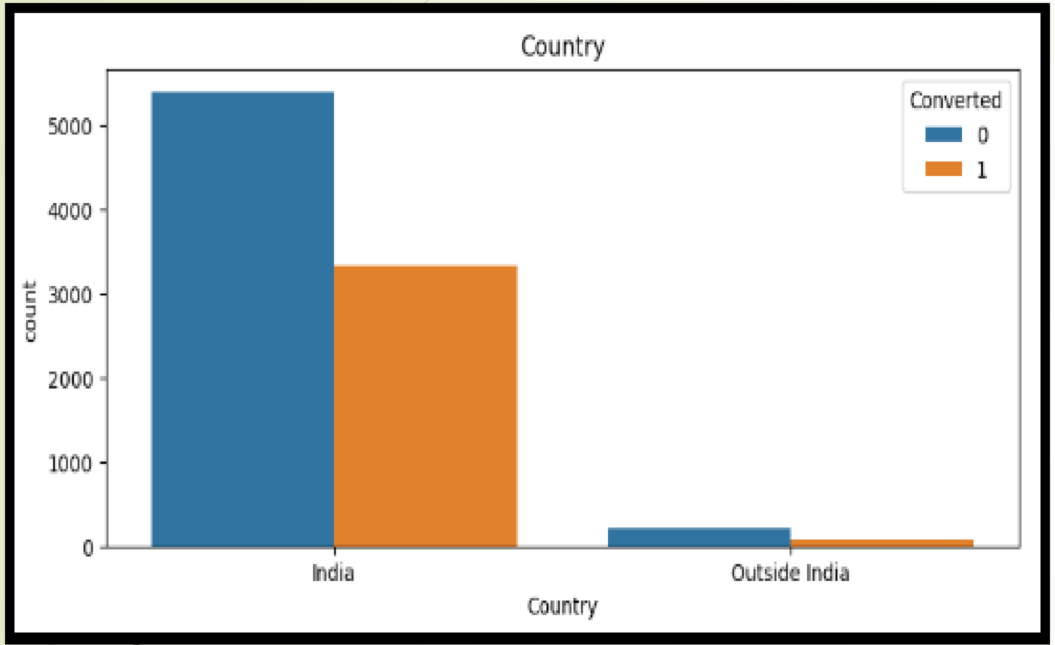


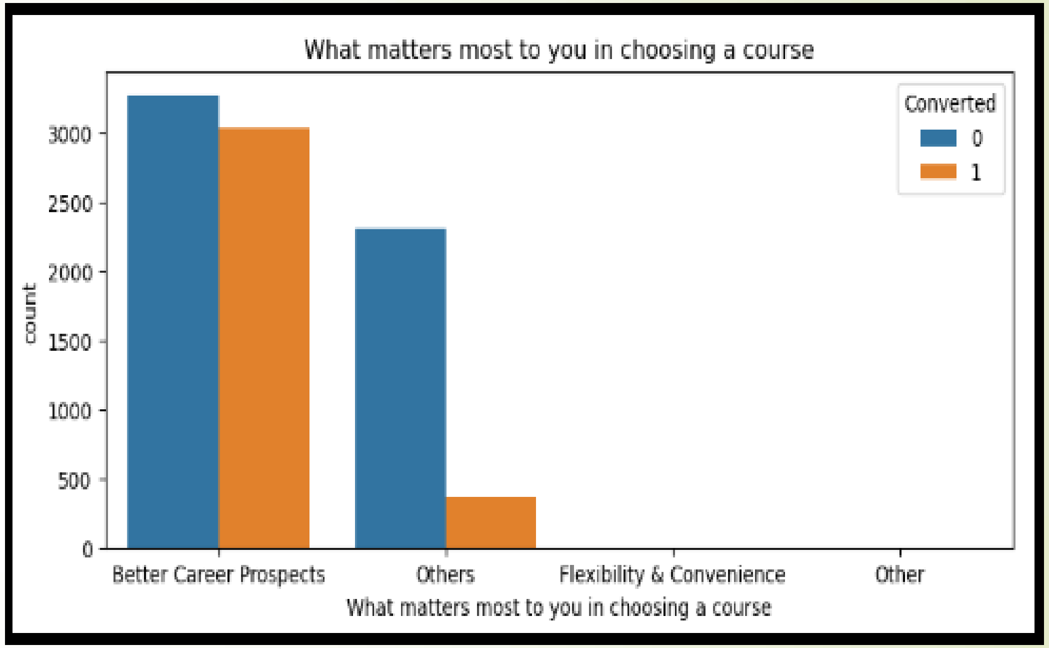


INSIGHTS:

- Leads originating from the 'Lead Add Form' have a significantly higher chance of being successfully converted.
- Leads with "Working Professionals" as their occupation have a higher likelihood of being successfully converted.

Visualizing Data & EDA : Categorical variables

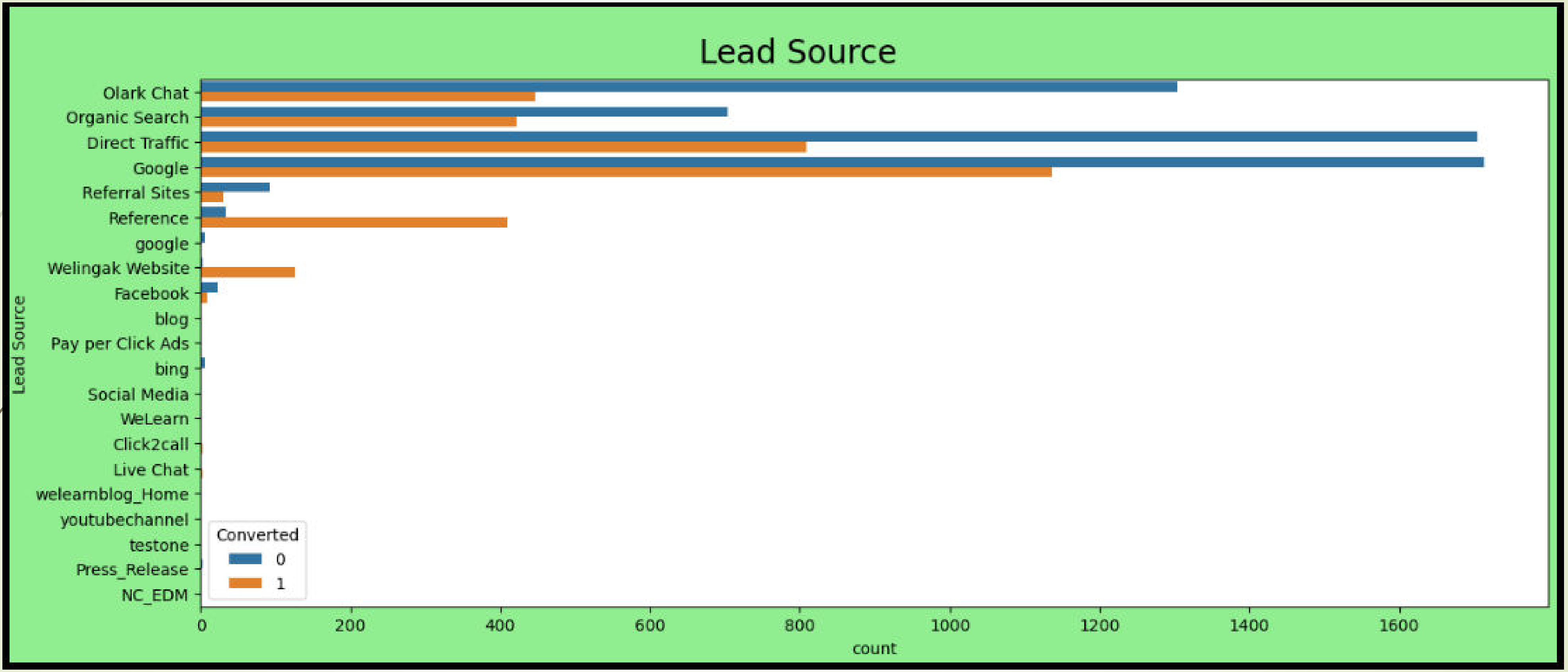




INSIGHTS:

- Mostly people are from India and looking for Better Career Prospects but most of them are not converted as they have may be financial issues or they can’t able to attend live courses, X Education can focus on diversified courses or they can launch some EMI option for conversions.

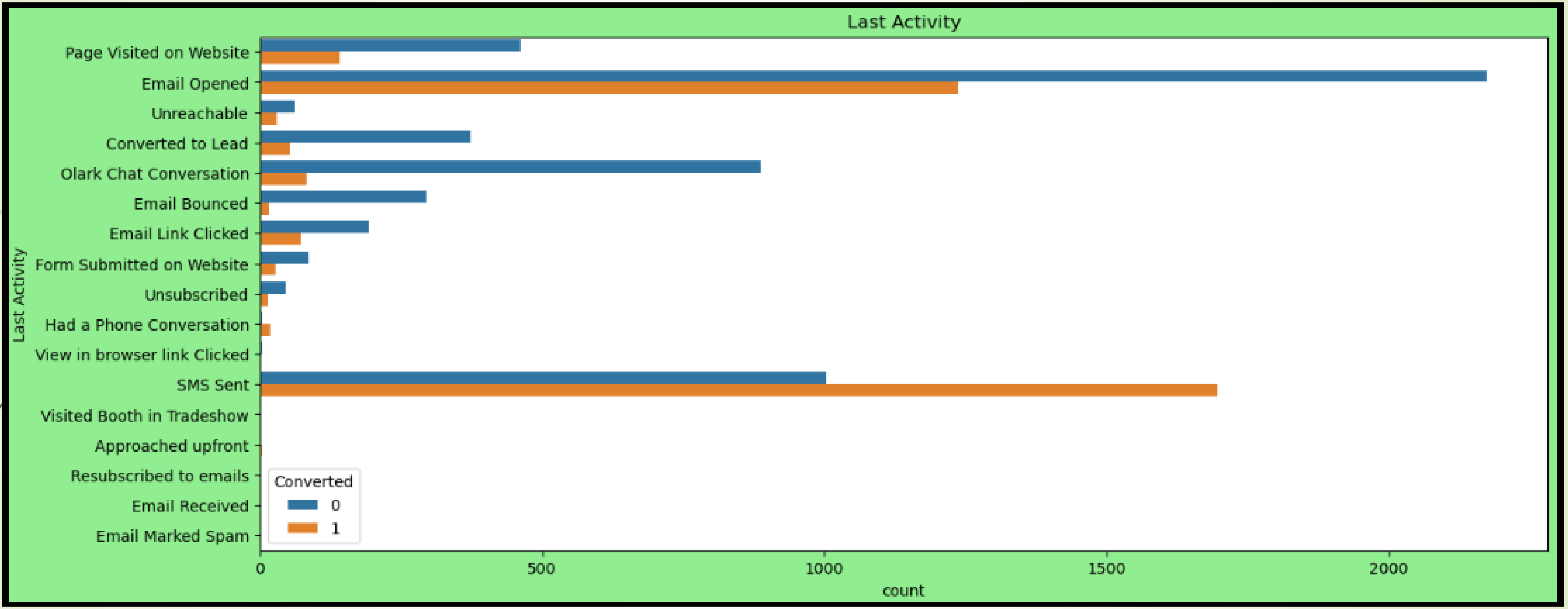
Visualizing Data & EDA : Categorical variables



INSIGHTS:

- Leads from the "Reference" type of Lead Source have a very high success rate, so the team should prioritize these customers. Although other sources bring in fewer leads, they still have a strong conversion rate. Additionally, customers coming through Organic Search also show a significantly higher chance of successful conversion.

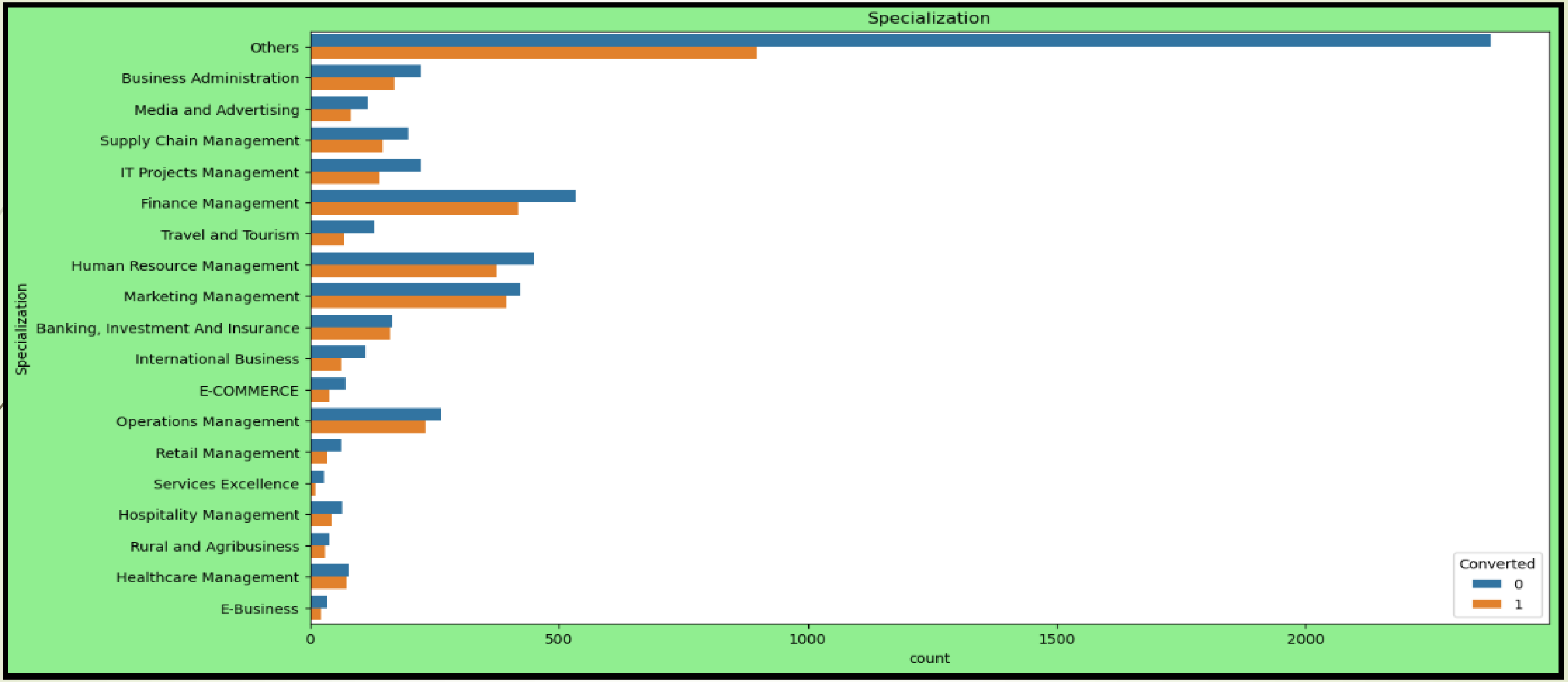
Visualizing Data & EDA : Categorical variables



INSIGHTS:

- The conversion rate of ‘SMS Sent’ have very high success ratio. Although other sources also have a strong conversion rate like E-mail opened. X Education can focus on this Activity and can increase their conversion rate.

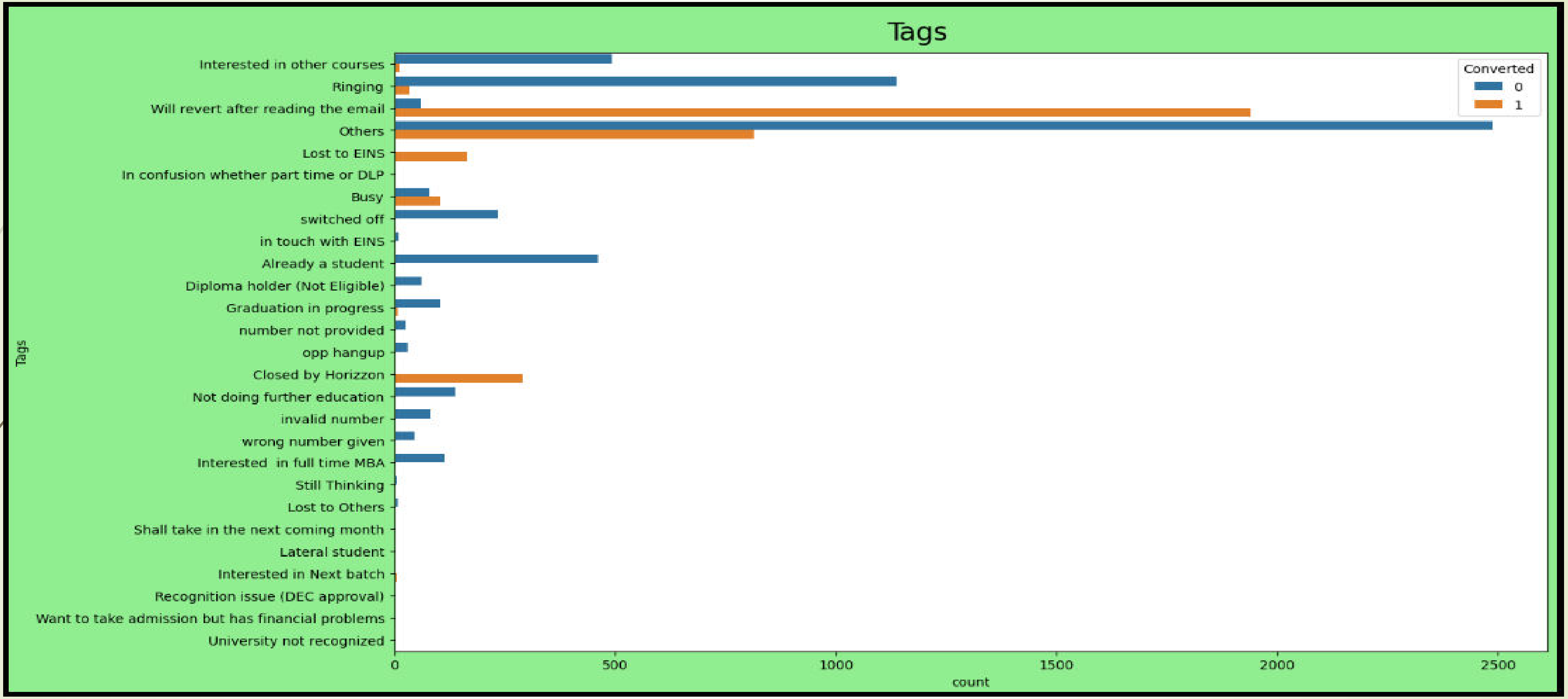
Visualizing Data & EDA : Categorical variables



INSIGHTS:

- People who have mentioned their Specialization in the form have higher chance of opting the course as compared to those who didn't mention.

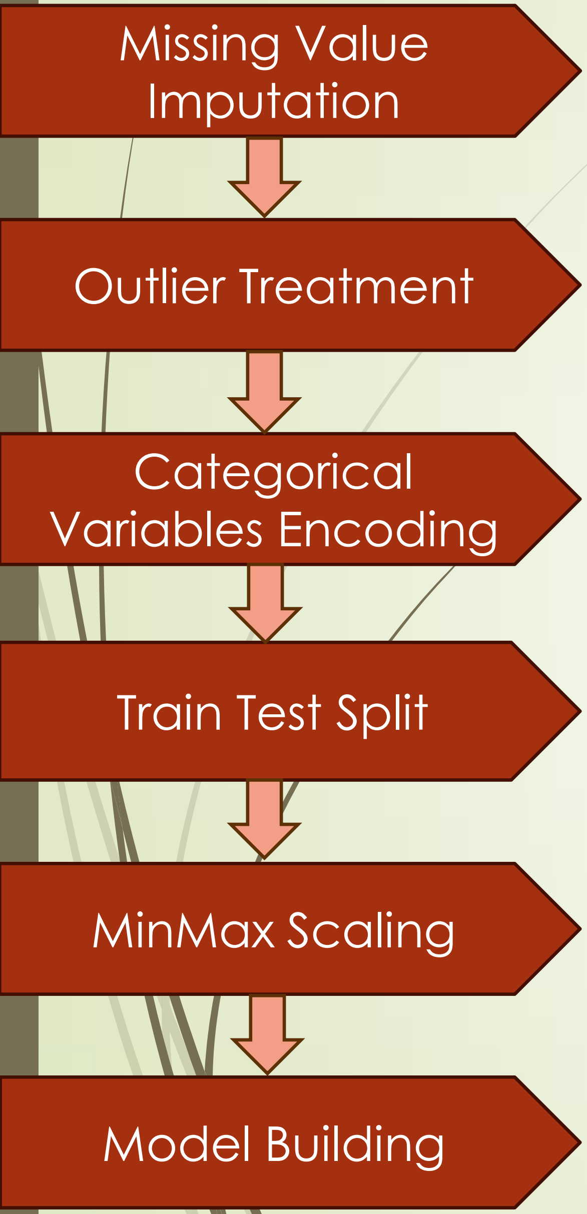
Visualizing Data & EDA : Categorical variables



INSIGHTS:

- People with a Tag ‘Will revert after reading the email’ have very high chance of conversion. Company can focus on creating and sending a good creative e-mail to the applicants which ensures to give more and more information regarding respective courses to the applicants.

Data Preparation



Missing Value Imputation

Missing value imputation done on the basis of business understanding by using mean, median for numerical columns and mode or creating other category for categorical columns.

Outlier Treatment

Identify outliers for numerical columns by plotting pairplot and boxplot and treated them on the basis of business understanding.

Categorical Variables Encoding

Binary variables that contain only 'Yes' or 'No' values have been replaced with 1 for 'Yes' and 0 for 'No’.
Dummy variables created for categorical columns that have 2 or more categories as their values.

Train Test Split

Data has been splitted into Train and Test in 70 : 30 ratio. The training set is used to build the model, while the test set is used to assess its performance.

MinMax Scaling

Performed MinMax Scaling (fit_transform) on Train dataset and MinMax Scaling (transform) on Test dataset for numerical variables only.

Model Building

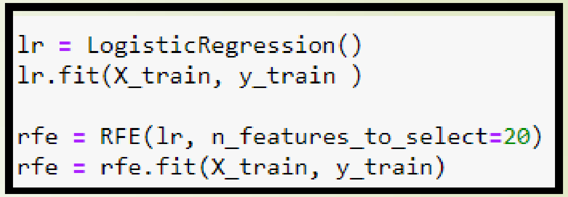
Creating first model with all the variables and then manual fine tune the model on the basis of high correlation, p-values and VIF.

Data Preparation

R F E

Recursive Feature Elimination

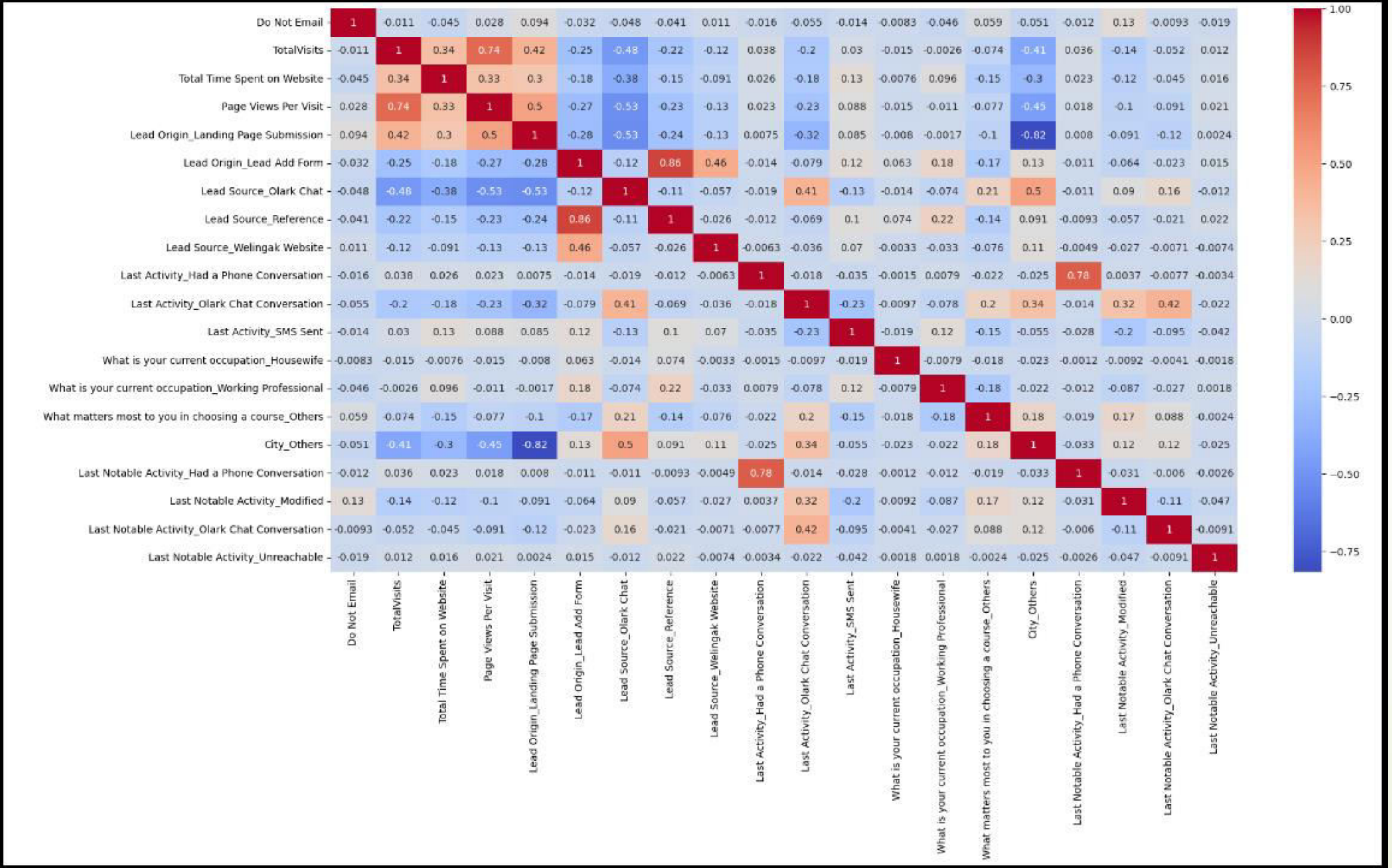
Imported and applied RFE to select top 20 variables for training model



```
lr = LogisticRegression()
lr.fit(X_train, y_train )

rfe = RFE(lr, n_features_to_select=20)
rfe = rfe.fit(X_train, y_train)
```

PAIRWISE CORRELATION



- It can be seen that 'Lead Origin_Lead Add Form' has very high correlation (0.86) with 'Lead Source_Reference', so we'll drop 'Lead Source_Reference' column.
- 'Lead_Origin_Landing Page Submission' has very high correlation with 'City_Others', so We'll drop 'City_Others' column.
- 'Last Activity_Had a Phone Conversation' has high correlation with 'Last Notable Activity_Had a Phone Conversation', so we'll drop 'Last Notable Activity_Had a Phone Conversation’.
- 'TotalVisits' has high correlation with 'Page Views Per Visit', so we'll drop 'Page Views Per Visit’ column.

MODEL BUILDING : Approach

- Recursive Feature Elimination (RFE) has been used to get top 20 features.
- After analyzing the pairplot, we removed four columns because they showed a high degree of correlation with other variables.
- Remaining 16 features are shown below:
  - Do Not
  - Email
  - TotalVisits
  - Total Time Spent on Website
  - Lead Origin_Landing Page Submission
  - Lead Origin_Lead Add Form
  - Lead Source_Olark Chat
  - Lead Source_Welingak Website
  - Last Activity_Had a Phone Conversation
  - Last Activity_Olark Chat Conversation
  - Last Activity_SMS Sent
  - What is your current occupation_Housewife
  - What is your current occupation_Working Professional
  - What matters most to you in choosing a course_Others
  - Last Notable Activity_Modified
  - Last Notable Activity_Olark Chat Conversation
  - Last Notable Activity_Unreachable

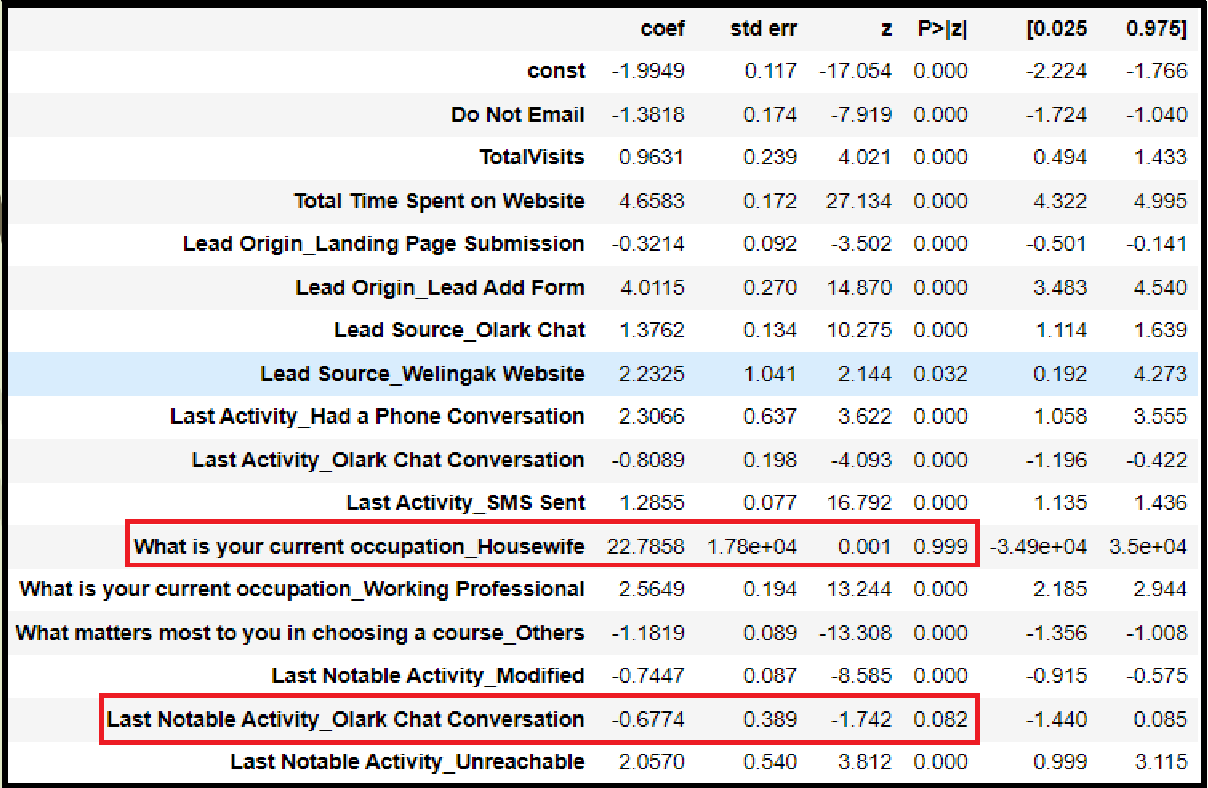
MODEL BUILDING : Approach

- We began developing the model by applying Logistic Regression using the Generalized Linear Model (GLM) approach in the statsmodels library, utilizing the remaining 16 features.
- We then manually fine-tuned the model to identify statistically significant features by reviewing the p-values, while also addressing multicollinearity by checking the Variance Inflation Factors (VIF). We retained features with p-values below 0.05 and VIF values under 5.
- A total of 3 models were developed, and after building each model, we reviewed the p-values of all beta coefficients and the VIFs. Features identified as problematic were removed in the subsequent model iteration. Additionally, we evaluated the overall model accuracy and confusion matrix after each new model to assess how its performance compared to the previous version.

MODEL BUILDING : Model -1

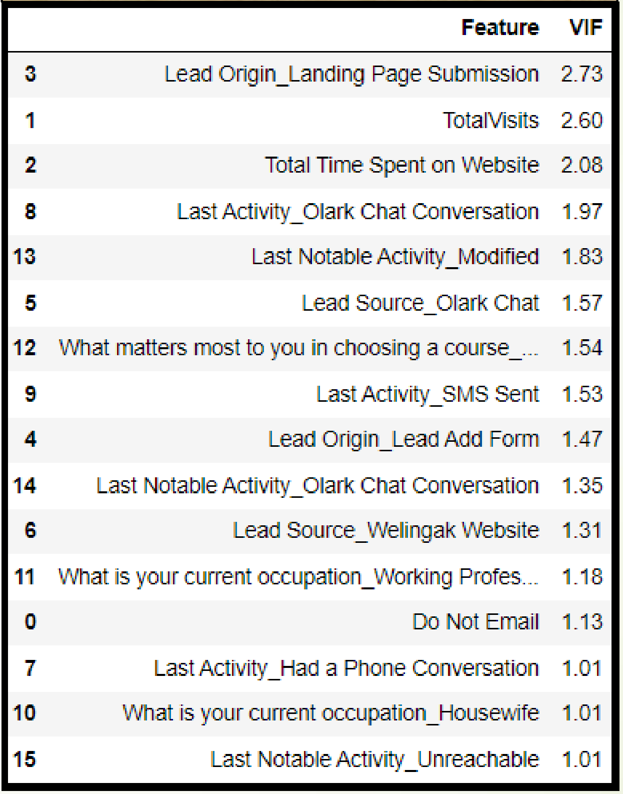
Logistic regression Model - 1 with 16 features :

Model Summary (Coefficients with p-value)



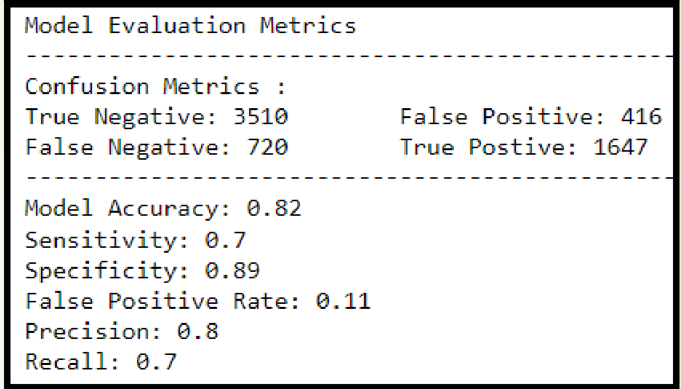
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9949	0.117	-17.054	0.000	-2.224	-1.766
Do Not Email	-1.3818	0.174	-7.919	0.000	-1.724	-1.040
TotalVisits	0.9631	0.239	4.021	0.000	0.494	1.433
Total Time Spent on Website	4.6583	0.172	27.134	0.000	4.322	4.995
Lead Origin_Landing Page Submission	-0.3214	0.092	-3.502	0.000	-0.501	-0.141
Lead Origin_Lead Add Form	4.0115	0.270	14.870	0.000	3.483	4.540
Lead Source_Olark Chat	1.3762	0.134	10.275	0.000	1.114	1.639
Lead Source_Welingak Website	2.2325	1.041	2.144	0.032	0.192	4.273
Last Activity_Had a Phone Conversation	2.3066	0.637	3.622	0.000	1.058	3.555
Last Activity_Olark Chat Conversation	-0.8089	0.198	-4.093	0.000	-1.196	-0.422
Last Activity_SMS Sent	1.2855	0.077	16.792	0.000	1.135	1.436
What is your current occupation_Housewife	22.7858	1.78e+04	0.001	0.999	-3.49e+04	3.5e+04
What is your current occupation_Working Professional	2.5649	0.194	13.244	0.000	2.185	2.944
What matters most to you in choosing a course_Others	-1.1819	0.089	-13.308	0.000	-1.356	-1.008
Last Notable Activity_Modified	-0.7447	0.087	-8.585	0.000	-0.915	-0.575
Last Notable Activity_Olark Chat Conversation	-0.6774	0.389	-1.742	0.082	-1.440	0.085
Last Notable Activity_Unreachable	2.0570	0.540	3.812	0.000	0.999	3.115

V I F



	Feature	VIF
3	Lead Origin_Landing Page Submission	2.73
1	TotalVisits	2.60
2	Total Time Spent on Website	2.08
8	Last Activity_Olark Chat Conversation	1.97
13	Last Notable Activity_Modified	1.83
5	Lead Source_Olark Chat	1.57
12	What matters most to you in choosing a course_...	1.54
9	Last Activity_SMS Sent	1.53
4	Lead Origin_Lead Add Form	1.47
14	Last Notable Activity_Olark Chat Conversation	1.35
6	Lead Source_Welingak Website	1.31
11	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
7	Last Activity_Had a Phone Conversation	1.01
10	What is your current occupation_Housewife	1.01
15	Last Notable Activity_Unreachable	1.01

Evaluation Metrics



```
Model Evaluation Metrics
-----------------------------------------------------
Confusion Metrics :
True Negative: 3510          False Positive: 416
False Negative: 720          True Postive: 1647
-----------------------------------------------------
Model Accuracy: 0.82
Sensitivity: 0.7
Specificity: 0.89
False Positive Rate: 0.11
Precision: 0.8
Recall: 0.7
```

Evaluation Metrics based on 0.5 cutoff probability. If probability is > 0.5 then 'Converted'=1(Yes) otherwise 0(No).

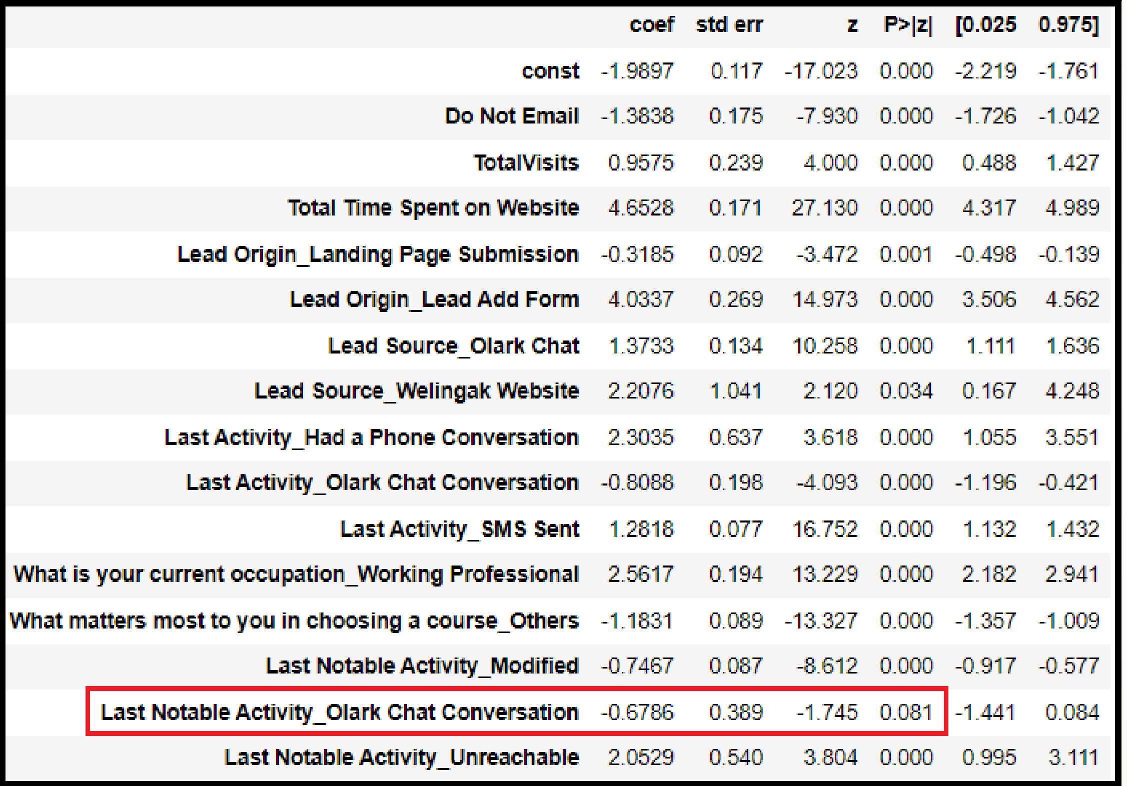
Observations :

- The features 'What is your current occupation_Housewife' and 'Last Notable Activity_Olark Chat Conversation' had p-values greater than 0.05, meaning their coefficients were not statistically significant. However, all the VIFs were within acceptable limits. As a result, in the next model, the 'What is your current occupation_Housewife' feature was removed from the set of predictors.

MODEL BUILDING : Model -2

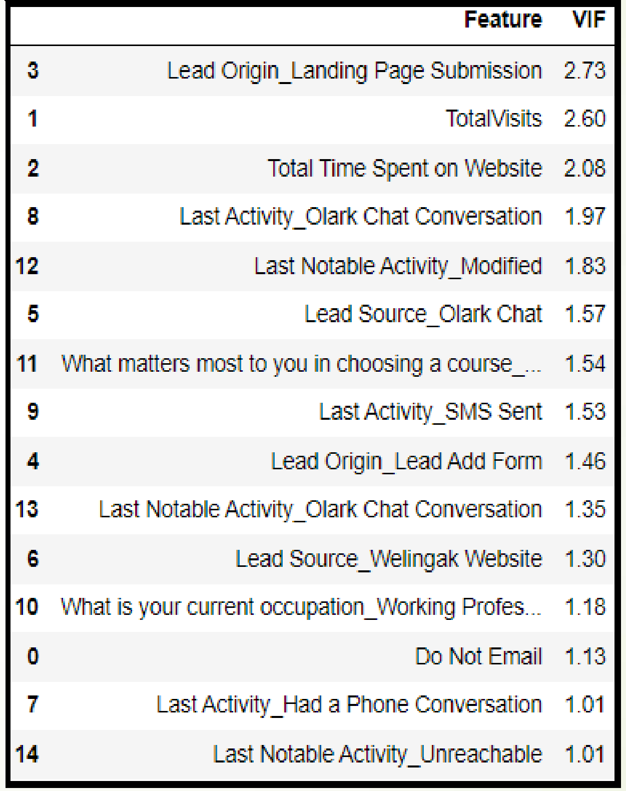
Logistic regression Model - 2 with 15 features :

Model Summary (Coefficients with p-value)



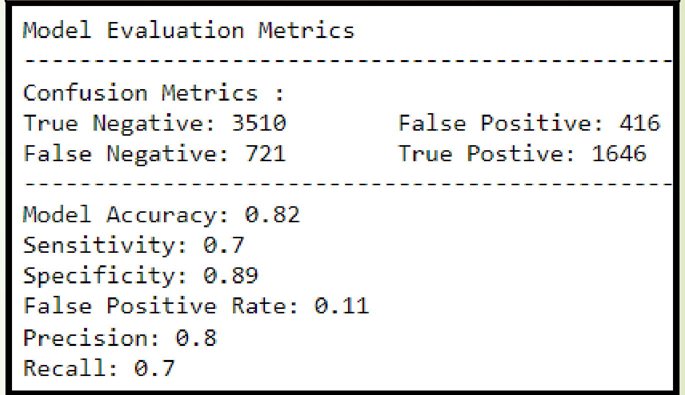
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9897	0.117	-17.023	0.000	-2.219	-1.761
Do Not Email	-1.3838	0.175	-7.930	0.000	-1.726	-1.042
TotalVisits	0.9575	0.239	4.000	0.000	0.488	1.427
Total Time Spent on Website	4.6528	0.171	27.130	0.000	4.317	4.989
Lead Origin_Landing Page Submission	-0.3185	0.092	-3.472	0.001	-0.498	-0.139
Lead Origin_Lead Add Form	4.0337	0.269	14.973	0.000	3.506	4.562
Lead Source_Olark Chat	1.3733	0.134	10.258	0.000	1.111	1.636
Lead Source_Welingak Website	2.2076	1.041	2.120	0.034	0.167	4.248
Last Activity_Had a Phone Conversation	2.3035	0.637	3.618	0.000	1.055	3.551
Last Activity_Olark Chat Conversation	-0.8088	0.198	-4.093	0.000	-1.196	-0.421
Last Activity_SMS Sent	1.2818	0.077	16.752	0.000	1.132	1.432
What is your current occupation_Working Professional	2.5617	0.194	13.229	0.000	2.182	2.941
What matters most to you in choosing a course_Others	-1.1831	0.089	-13.327	0.000	-1.357	-1.009
Last Notable Activity_Modified	-0.7467	0.087	-8.612	0.000	-0.917	-0.577
Last Notable Activity_Olark Chat Conversation	-0.6786	0.389	-1.745	0.081	-1.441	0.084
Last Notable Activity_Unreachable	2.0529	0.540	3.804	0.000	0.995	3.111

V I F



	Feature	VIF
3	Lead Origin_Landing Page Submission	2.73
1	TotalVisits	2.60
2	Total Time Spent on Website	2.08
8	Last Activity_Olark Chat Conversation	1.97
12	Last Notable Activity_Modified	1.83
5	Lead Source_Olark Chat	1.57
11	What matters most to you in choosing a course_...	1.54
9	Last Activity_SMS Sent	1.53
4	Lead Origin_Lead Add Form	1.46
13	Last Notable Activity_Olark Chat Conversation	1.35
6	Lead Source_Welingak Website	1.30
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
7	Last Activity_Had a Phone Conversation	1.01
14	Last Notable Activity_Unreachable	1.01

Evaluation Metrics



```
Model Evaluation Metrics
--------------------------------------------------------
Confusion Metrics :
True Negative: 3510          False Positive: 416
False Negative: 721          True Postive: 1646
--------------------------------------------------------
Model Accuracy: 0.82
Sensitivity: 0.7
Specificity: 0.89
False Positive Rate: 0.11
Precision: 0.8
Recall: 0.7
```

Evaluation Metrics based on 0.5 cutoff probability. If probability is > 0.5 then ‘Converted’=1(Yes) otherwise 0(No).

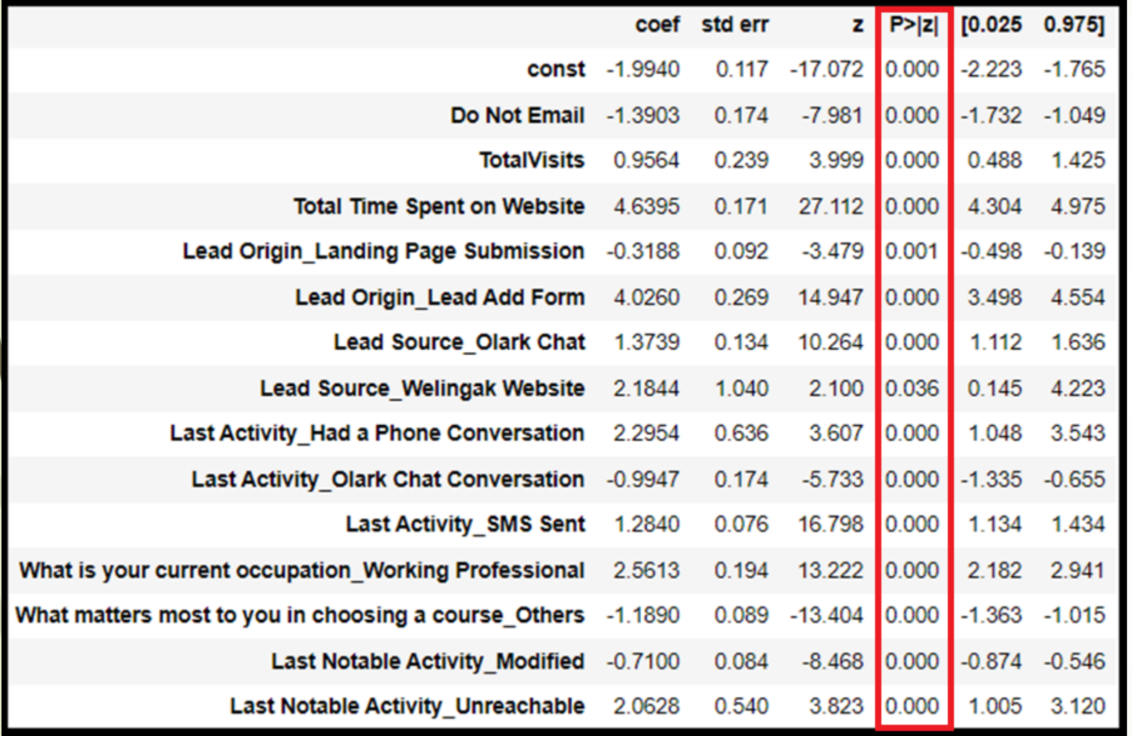
Observations :

- The feature 'Last Notable Activity_Olark Chat Conversation' had p-values greater than 0.05, meaning their coefficients were not statistically significant. However, all the VIFs were within acceptable limits. As a result, in the next model, 'Last Notable Activity_Olark Chat Conversation' feature was removed from the set of predictors.

MODEL BUILDING : Model -3

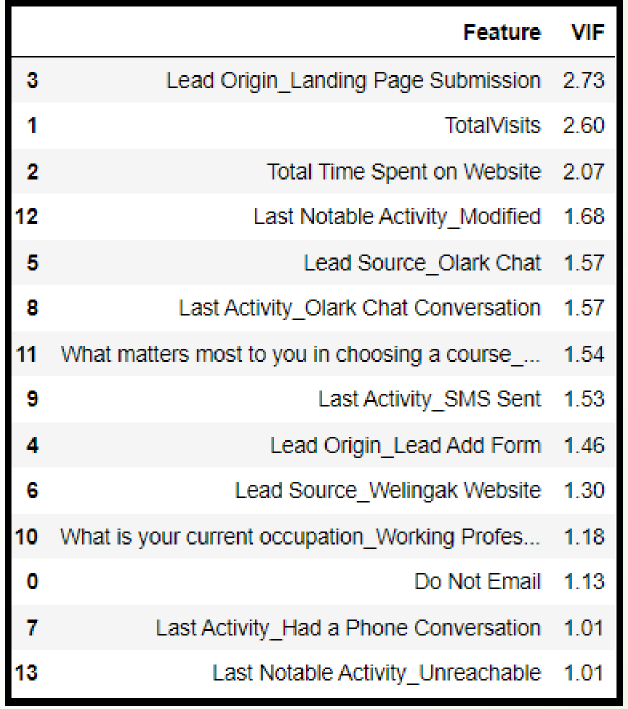
Logistic regression Model - 3 with 14 features :

Model Summary (Coefficients with p-value)



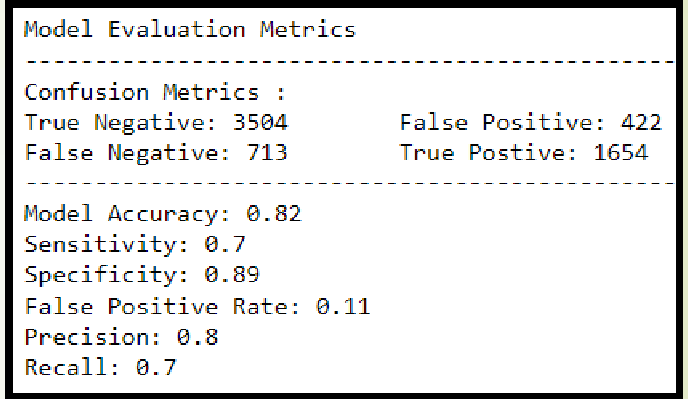
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9940	0.117	-17.072	0.000	-2.223	-1.765
Do Not Email	-1.3903	0.174	-7.981	0.000	-1.732	-1.049
TotalVisits	0.9564	0.239	3.999	0.000	0.488	1.425
Total Time Spent on Website	4.6395	0.171	27.112	0.000	4.304	4.975
Lead Origin_Landing Page Submission	-0.3188	0.092	-3.479	0.001	-0.498	-0.139
Lead Origin_Lead Add Form	4.0260	0.269	14.947	0.000	3.498	4.554
Lead Source_Olark Chat	1.3739	0.134	10.264	0.000	1.112	1.636
Lead Source_Welingak Website	2.1844	1.040	2.100	0.036	0.145	4.223
Last Activity_Had a Phone Conversation	2.2954	0.636	3.607	0.000	1.048	3.543
Last Activity_Olark Chat Conversation	-0.9947	0.174	-5.733	0.000	-1.335	-0.655
Last Activity_SMS Sent	1.2840	0.076	16.798	0.000	1.134	1.434
What is your current occupation_Working Professional	2.5613	0.194	13.222	0.000	2.182	2.941
What matters most to you in choosing a course_Others	-1.1890	0.089	-13.404	0.000	-1.363	-1.015
Last Notable Activity_Modified	-0.7100	0.084	-8.468	0.000	-0.874	-0.546
Last Notable Activity_Unreachable	2.0628	0.540	3.823	0.000	1.005	3.120

V I F



	Feature	VIF
3	Lead Origin_Landing Page Submission	2.73
1	TotalVisits	2.60
2	Total Time Spent on Website	2.07
12	Last Notable Activity_Modified	1.68
5	Lead Source_Olark Chat	1.57
8	Last Activity_Olark Chat Conversation	1.57
11	What matters most to you in choosing a course_...	1.54
9	Last Activity_SMS Sent	1.53
4	Lead Origin_Lead Add Form	1.46
6	Lead Source_Welingak Website	1.30
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
7	Last Activity_Had a Phone Conversation	1.01
13	Last Notable Activity_Unreachable	1.01

Evaluation Metrics



```
Model Evaluation Metrics
-------------------------------------------------------
Confusion Metrics :
True Negative: 3504          False Positive: 422
False Negative: 713          True Postive: 1654
-------------------------------------------------------
Model Accuracy: 0.82
Sensitivity: 0.7
Specificity: 0.89
False Positive Rate: 0.11
Precision: 0.8
Recall: 0.7
```

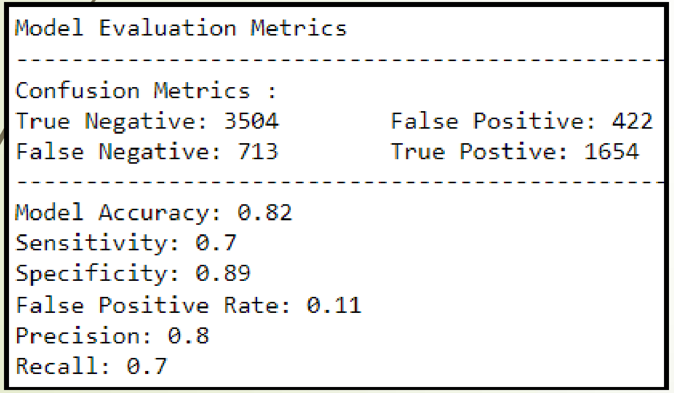
Evaluation Metrics based on 0.5 cutoff probability. If probability is > 0.5 then 'Converted'=1(Yes) otherwise 0(No).

Observations :

- We can see that all the beta coefficients are now statistically significant also there is no multicollinearity present. Also the model accuracy is 0.82, sensitivity is 0.7 and specificity is 0.89, So we choose lr_3 as our final model for Predictions.

Prediction & Model Evaluation On Training Data – Cutoff is 0.5

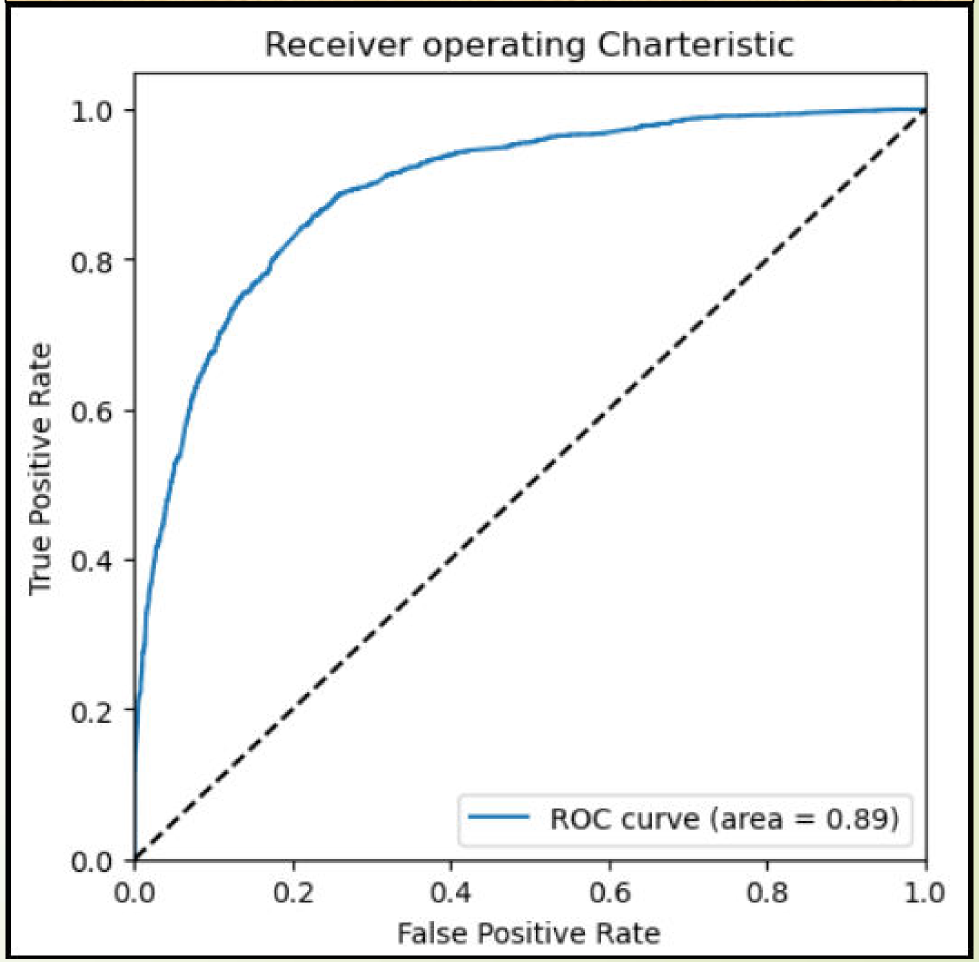
- Using Model 3, we estimated the probability for each observation in the training dataset and applied a cutoff of 0.5 to classify the target variable 'Converted.' If the probability was above 0.5, 'Converted' was set to 1 (Yes); otherwise, it was set to 0 (No).
- After predicting the target on our training data set, we calculated evaluation metrics as below:



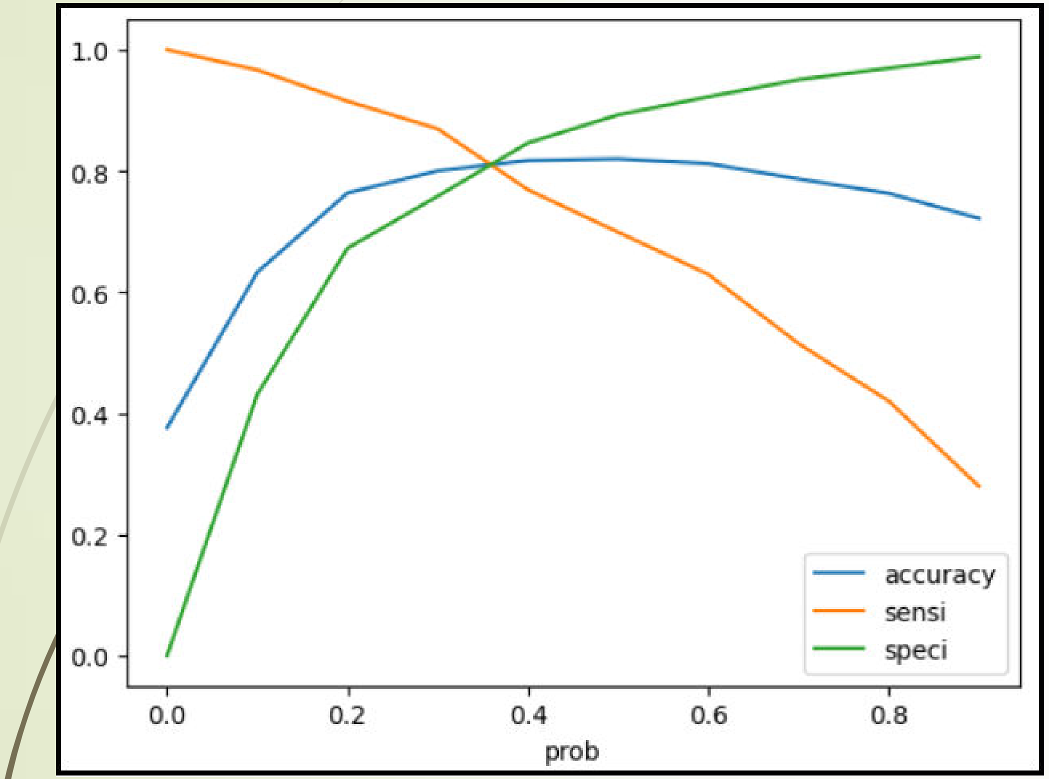
```
Model Evaluation Metrics
--------------------------------------------------
Confusion Metrics :
True Negative: 3504          False Positive: 422
False Negative: 713          True Postive: 1654
--------------------------------------------------
Model Accuracy: 0.82
Sensitivity: 0.7
Specificity: 0.89
False Positive Rate: 0.11
Precision: 0.8
Recall: 0.7
```

- Our model has bad Sensitivity (with probability cut-off .5). By doing trade-off between Sensitivity-Specificity optimal probability cut-off value has been calculated.

ROC Curve

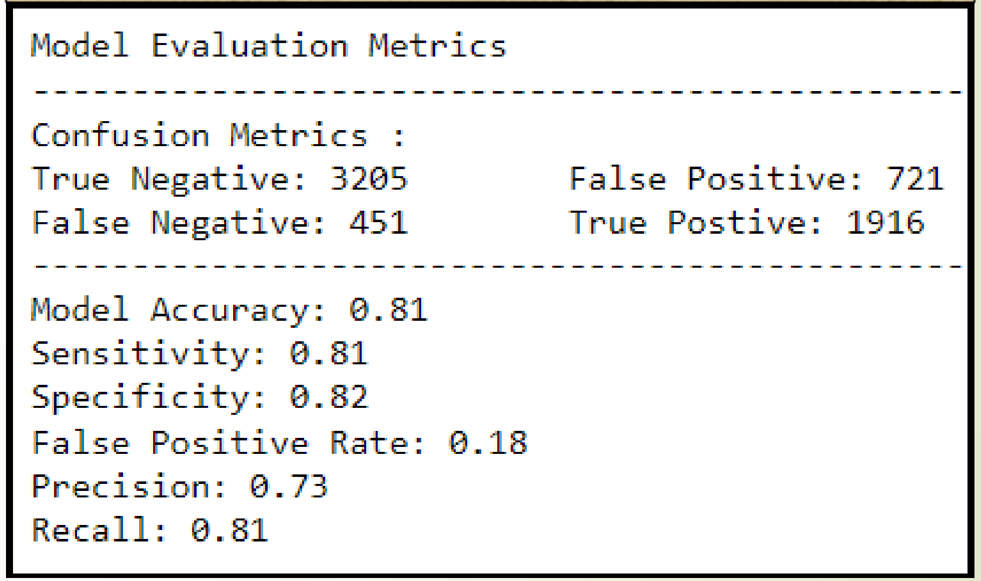


Finding Optimal Probability Cutoff & Evaluating on Train Data



- In above plot, it's visible that 0.35 is the optimal point to set as cutoff probability for our model.

Evaluation Metrics based on 0.35 cutoff probability. If probability is > 0.35 then 'Converted'=1(Yes) otherwise 0(No).



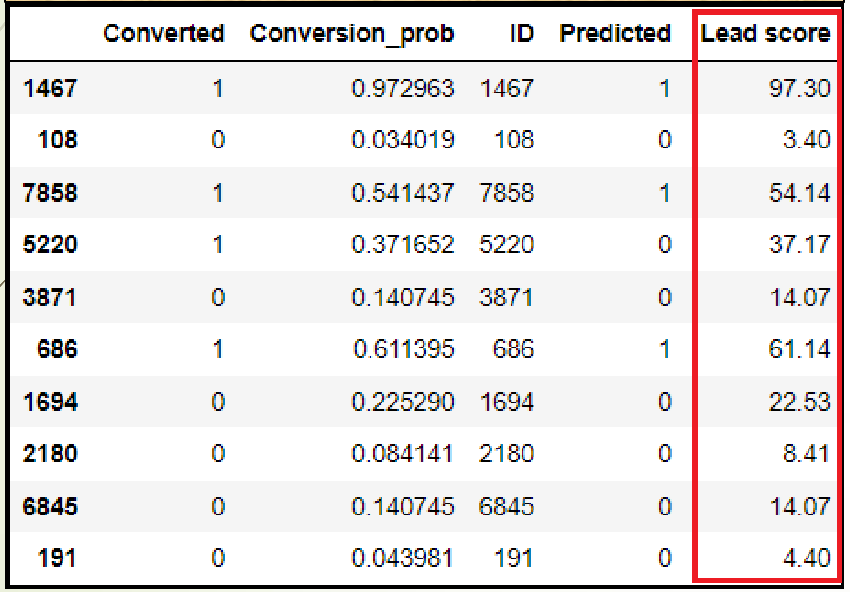
```
Model Evaluation Metrics
---------------------------------------------------
Confusion Metrics :
True Negative: 3205          False Positive: 721
False Negative: 451          True Postive: 1916
---------------------------------------------------
Model Accuracy: 0.81
Sensitivity: 0.81
Specificity: 0.82
False Positive Rate: 0.18
Precision: 0.73
Recall: 0.81
```

Observation :

- The sensitivity of our model has improved without significantly lowering overall accuracy. The new specificity is also within an acceptable range.

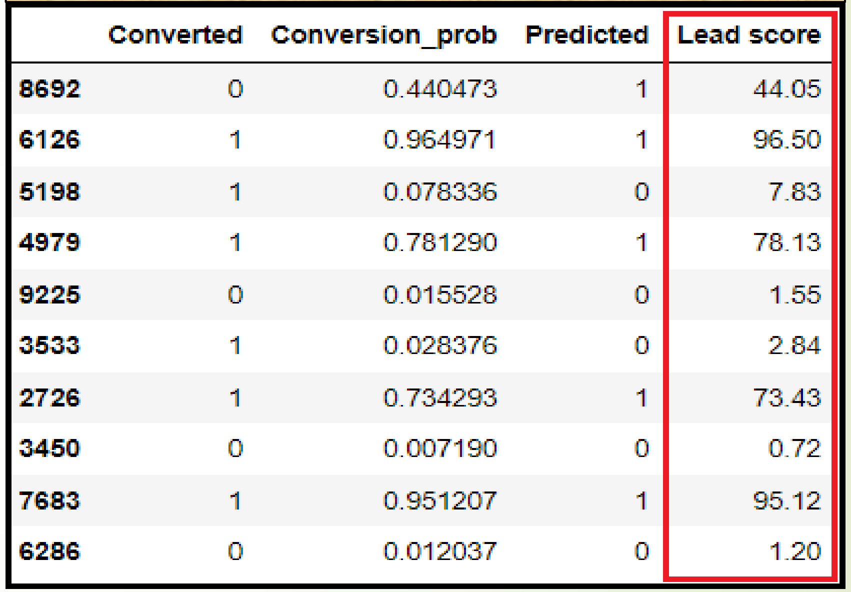
Prediction & Generating Lead Score (Business Requirement)

Generating Lead Score on Train Data



	Converted	Conversion_prob	ID	Predicted	Lead score
1467	1	0.972963	1467	1	97.30
108	0	0.034019	108	0	3.40
7858	1	0.541437	7858	1	54.14
5220	1	0.371652	5220	0	37.17
3871	0	0.140745	3871	0	14.07
686	1	0.611395	686	1	61.14
1694	0	0.225290	1694	0	22.53
2180	0	0.084141	2180	0	8.41
6845	0	0.140745	6845	0	14.07
191	0	0.043981	191	0	4.40

Generating Lead Score on Test Data

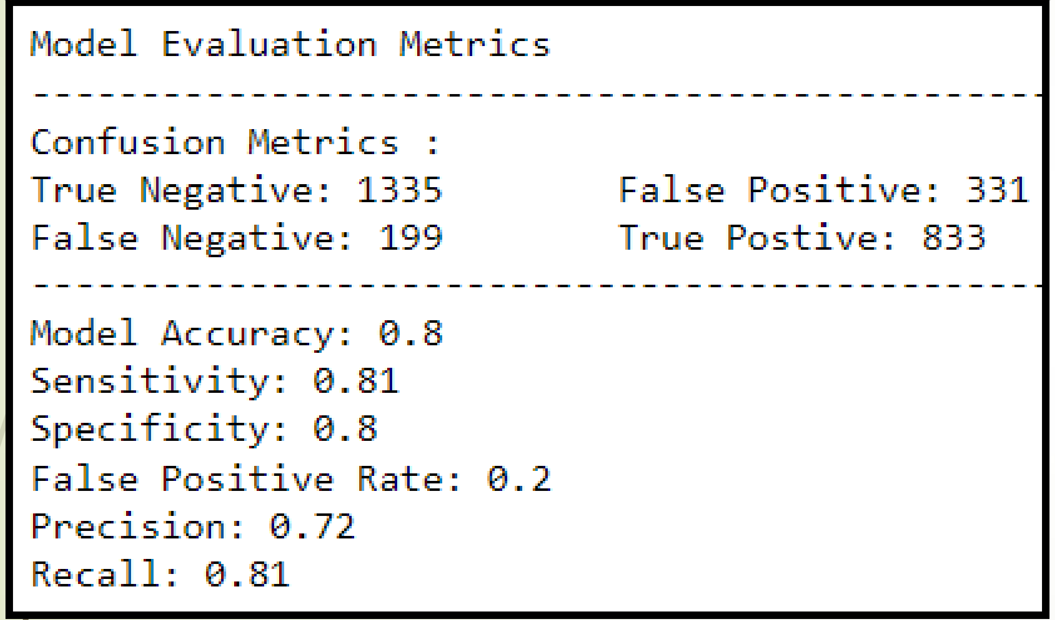


	Converted	Conversion_prob	Predicted	Lead score
8692	0	0.440473	1	44.05
6126	1	0.964971	1	96.50
5198	1	0.078336	0	7.83
4979	1	0.781290	1	78.13
9225	0	0.015528	0	1.55
3533	1	0.028376	0	2.84
2726	1	0.734293	1	73.43
3450	0	0.007190	0	0.72
7683	1	0.951207	1	95.12
6286	0	0.012037	0	1.20

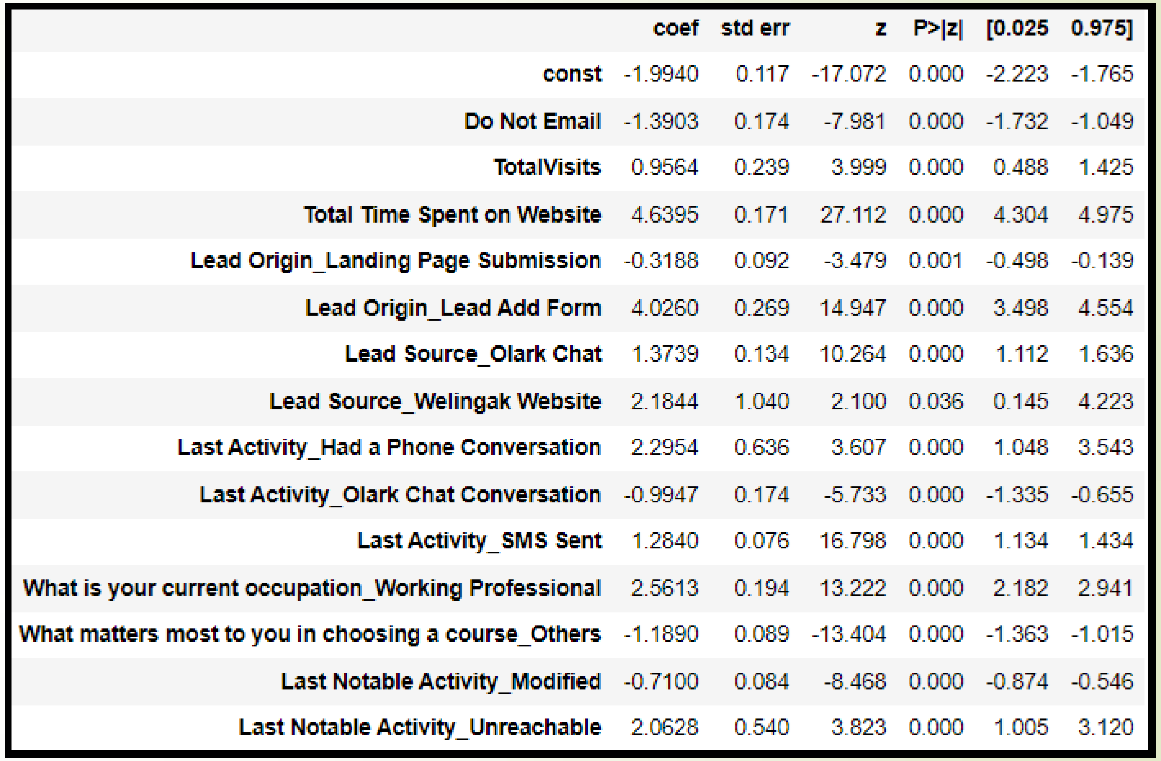
- Using Model 3 we calculated the probability on Test dataset and used cutoff =0.35 to predict the Predicted(0,1). As per business requirement we have created a column Lead Score (between 0 to 100) of the leads. A higher score means hot lead (most likely to convert), lower score implies cold lead (less likely to convert). We have multiplied the Conversion_prob with 100 to generate the Lead Score.

Model Evaluation on Test Data & Interpretation

Model Evaluation Metrics on Test dataset with probability cutoff of 0.35



```
Model Evaluation Metrics
------------------------------------------------
Confusion Metrics :
True Negative: 1335          False Positive: 331
False Negative: 199          True Postive: 833
------------------------------------------------
Model Accuracy: 0.8
Sensitivity: 0.81
Specificity: 0.8
False Positive Rate: 0.2
Precision: 0.72
Recall: 0.81
```



	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9940	0.117	-17.072	0.000	-2.223	-1.765
Do Not Email	-1.3903	0.174	-7.981	0.000	-1.732	-1.049
TotalVisits	0.9564	0.239	3.999	0.000	0.488	1.425
Total Time Spent on Website	4.6395	0.171	27.112	0.000	4.304	4.975
Lead Origin_Landing Page Submission	-0.3188	0.092	-3.479	0.001	-0.498	-0.139
Lead Origin_Lead Add Form	4.0260	0.269	14.947	0.000	3.498	4.554
Lead Source_Olark Chat	1.3739	0.134	10.264	0.000	1.112	1.636
Lead Source_Welingak Website	2.1844	1.040	2.100	0.036	0.145	4.223
Last Activity_Had a Phone Conversation	2.2954	0.636	3.607	0.000	1.048	3.543
Last Activity_Olark Chat Conversation	-0.9947	0.174	-5.733	0.000	-1.335	-0.655
Last Activity_SMS Sent	1.2840	0.076	16.798	0.000	1.134	1.434
What is your current occupation_Working Professional	2.5613	0.194	13.222	0.000	2.182	2.941
What matters most to you in choosing a course_Others	-1.1890	0.089	-13.404	0.000	-1.363	-1.015
Last Notable Activity_Modified	-0.7100	0.084	-8.468	0.000	-0.874	-0.546
Last Notable Activity_Unreachable	2.0628	0.540	3.823	0.000	1.005	3.120

- Model is performing well on test data with Sensitivity= 81%, Specificity= 80% and overall accuracy: 80%.
- Top 3 variables which contribute most towards the probability of a lead getting converted:
  - Total time spent on website
  - Lead Origin_Lead Add Form
  - What is your current occupation_Working Professional

CONCLUSION & RECOMMENDATION

- In line with business requirements, we calculated the Lead Score (ranging from 0 to 100) using the Logistic Regression model. A higher score indicates a hot lead (more likely to convert), while a lower score signifies a cold lead (less likely to convert).
- The Lead Score will help identify hot leads more quickly and efficiently, leading to a **reduction in lead conversion time** and an **increase in the conversion rate**. Leads should be ranked in descending order based on their Lead Scores.
- Leads with higher Lead Scores should be prioritized for phone calls or contact. These hot leads should receive special attention, such as assigning a dedicated Support Point of Contact (SPOC) to a small group of high-scoring leads, as they have a strong likelihood of conversion.
- Leads with medium Lead Scores are also promising candidates for conversion. They should be contacted, and the right questions should be asked to better understand their needs and challenges. Addressing their concerns, such as making changes to existing courses, introducing new courses, adjusting class schedules, or offering flexible financial options for fees, could help convert these leads successfully.
- Cold leads should be contacted once the business achieves a strong conversion rate with leads that have high and medium Lead Scores. Since the likelihood of conversion is lower with these leads, they can be included in the company’s more aggressive marketing strategy.

THANKYOU