Lead Scoring Case Study using logistic regression

SUBMITTED BY :
1. Deeksha Pant 2. Pankti Patel 3. Piyush Verma

Contents

Problem Statement

- X Education sells online courses to industry professionals.
- 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.

Business Objective

- Lead X wants us to build a model to give every lead a lead score between 0 - 100 . So that they can identify the Hot leads and increase their conversion rate as well.

- The CEO want to achieve a lead conversion rate of 80%.

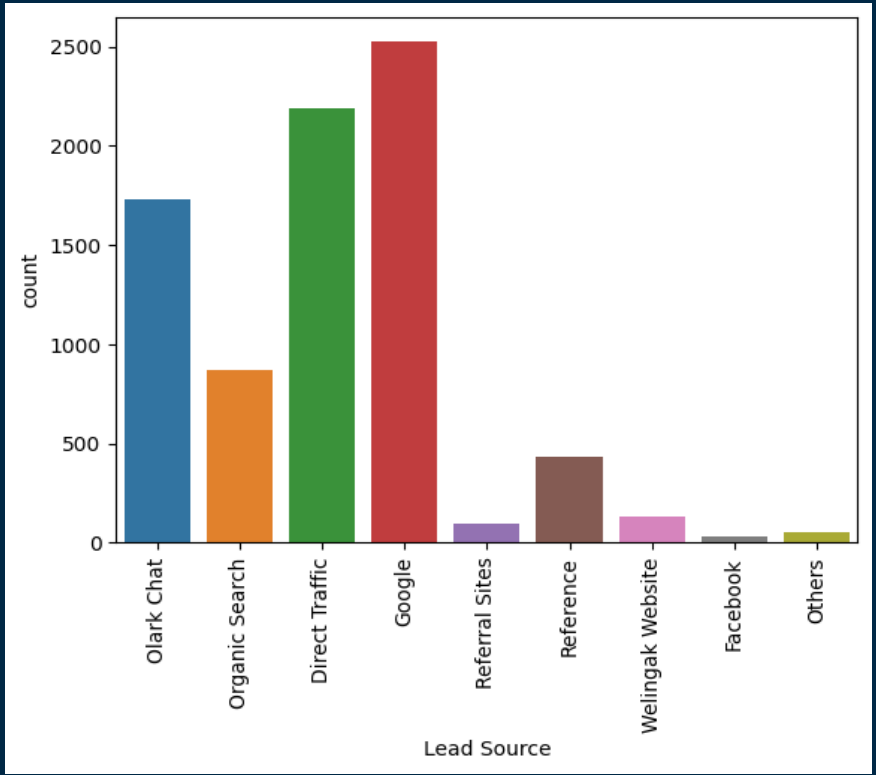
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

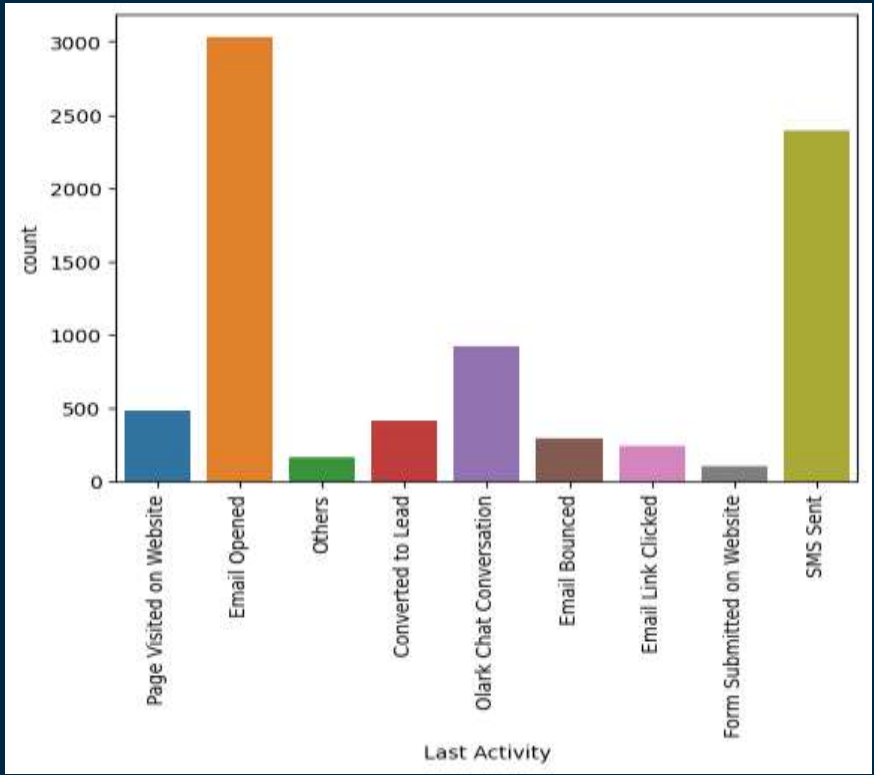
Solution Methodology

- Data cleaning and data Inspection.
- EDA
- Dummy Variable Creation
- Test-Train split
- Feature scaling
- Dropping highly correlated dummy variables
- Model Building (RFE Rsquared VIF and pvalues)
- Model Evaluation
- Checking Accuracy
- Finding Optimal Cutoff Point
- Making predictions on test set

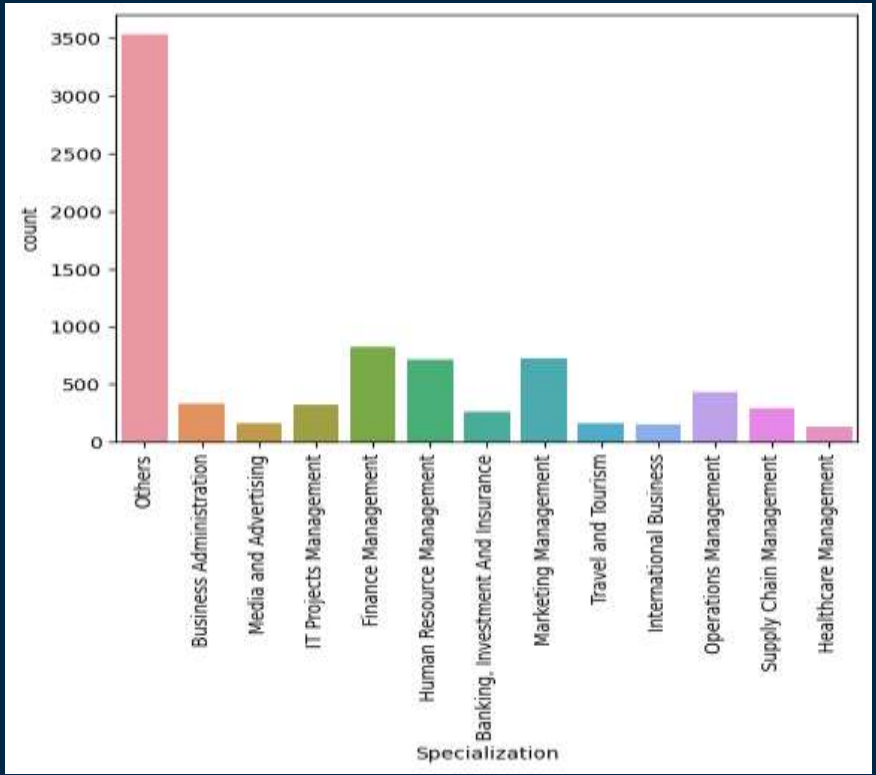
Exploratory Data Analysis

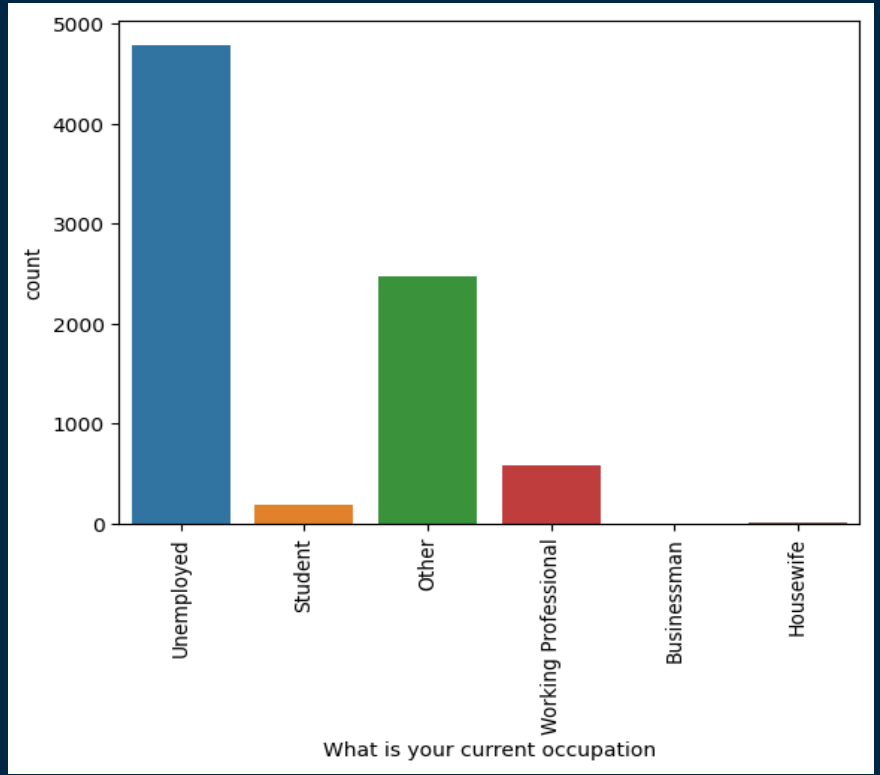
Univariate Analysis



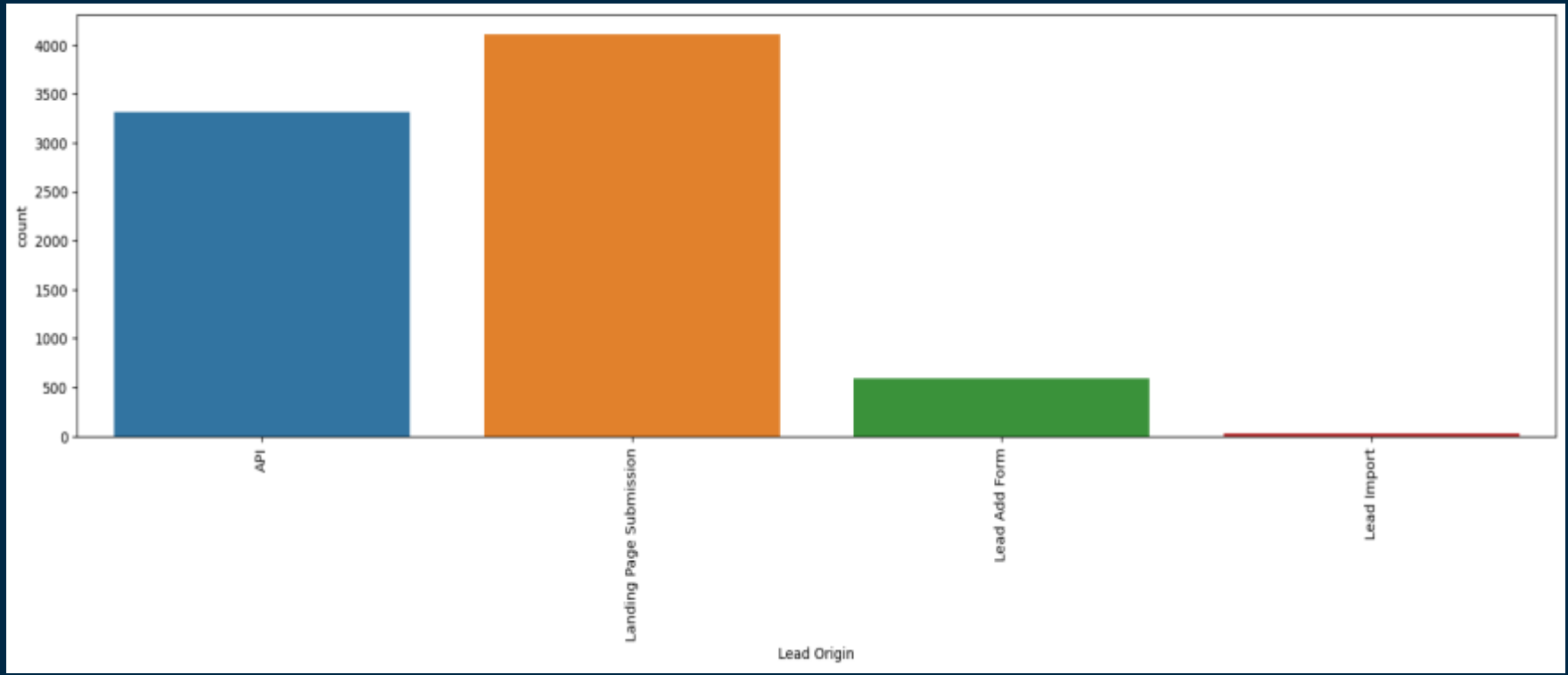


Univariate Analysis





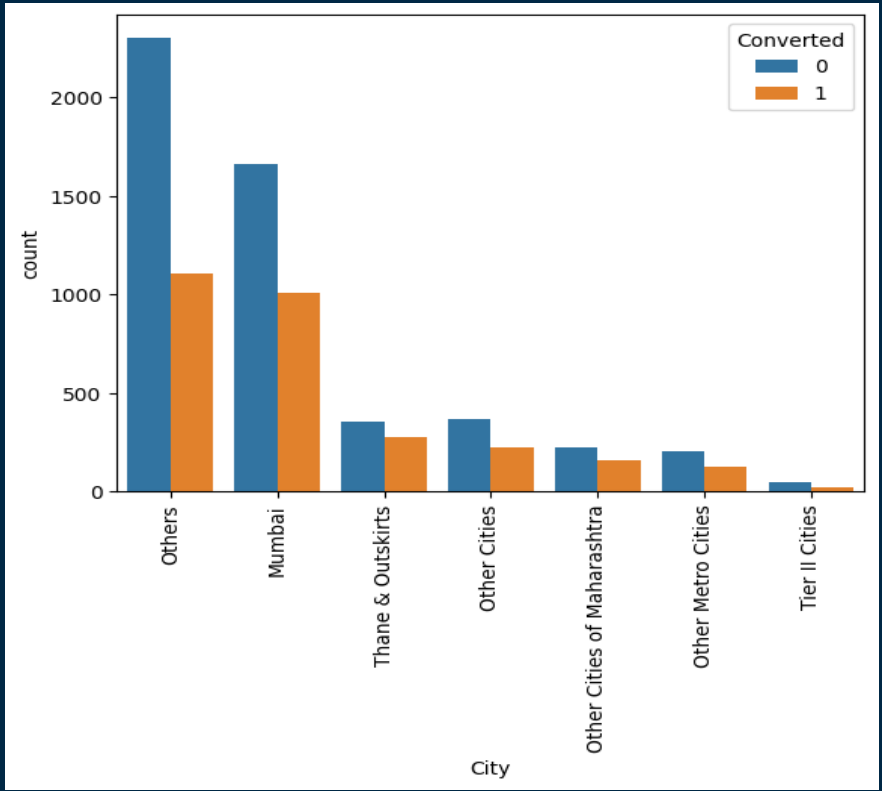
Univariate Analysis

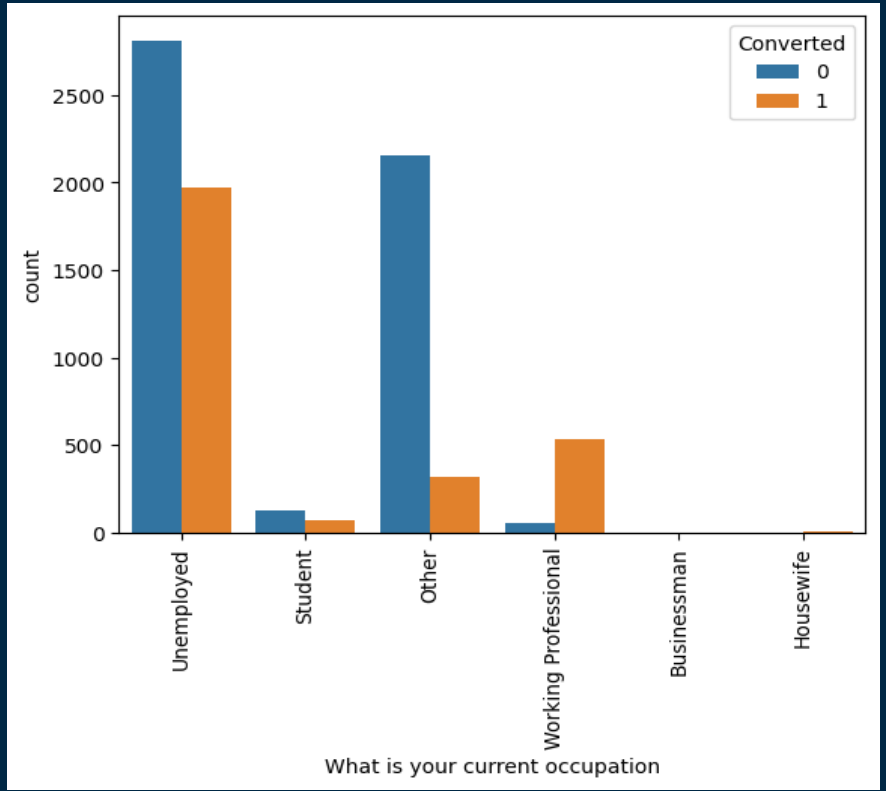


Univariate Analysis

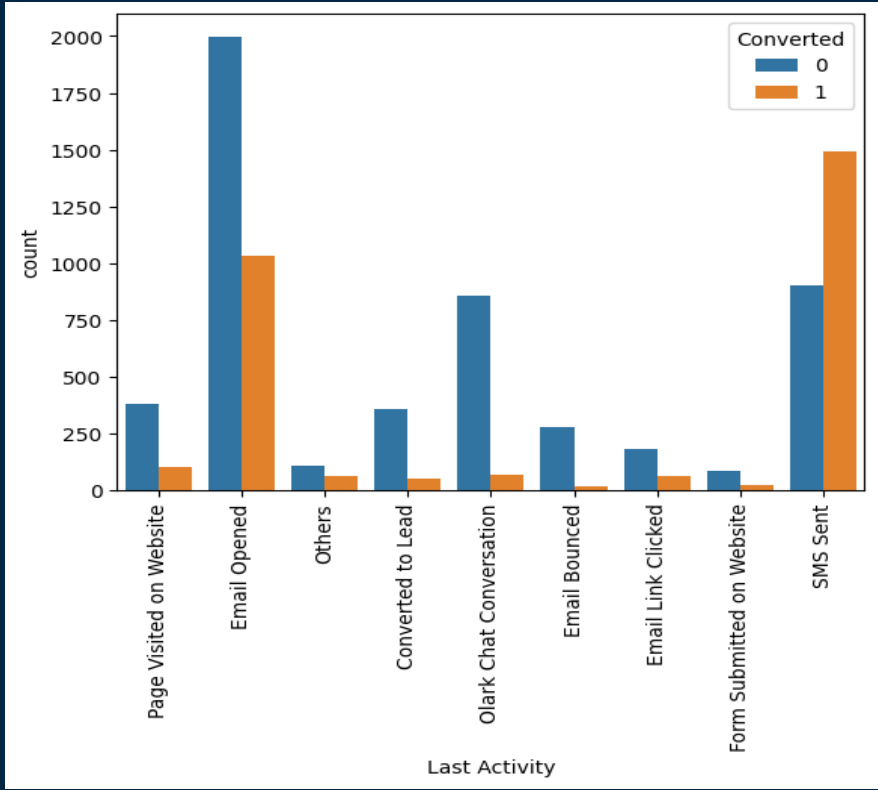
- Majority people are using either Google or Direct Traffic as lead source
- Unemployed are the majority of people who are visiting the site
- The last activity for majority of the leads is Email opened
- Majority leads have Landing Page submission as the Lead Origin

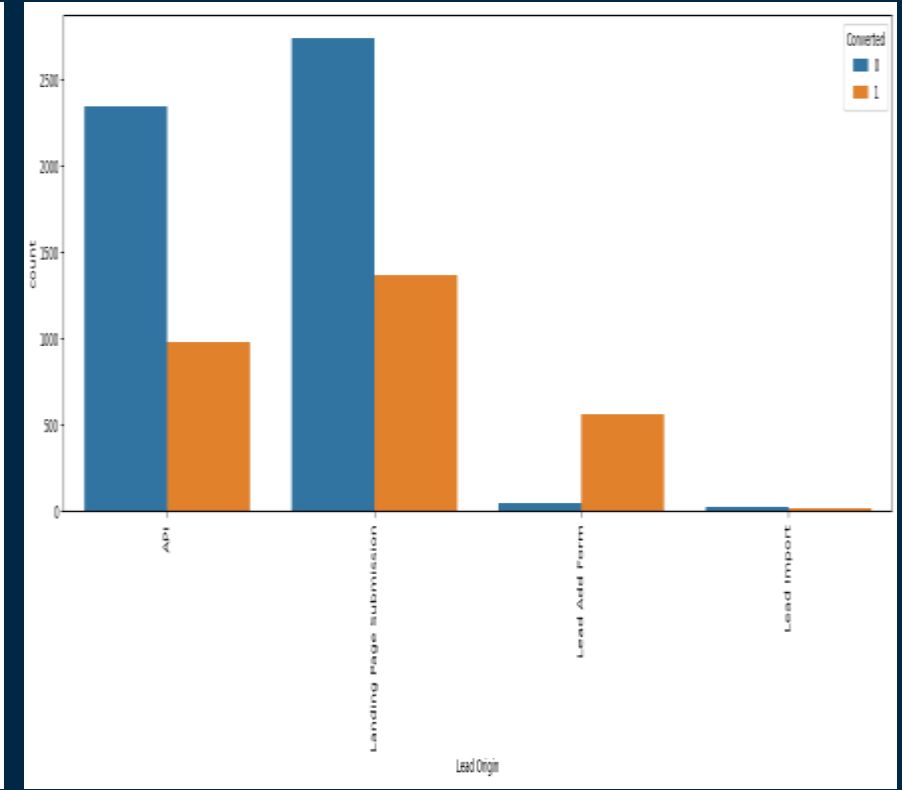
Bivariate Analysis



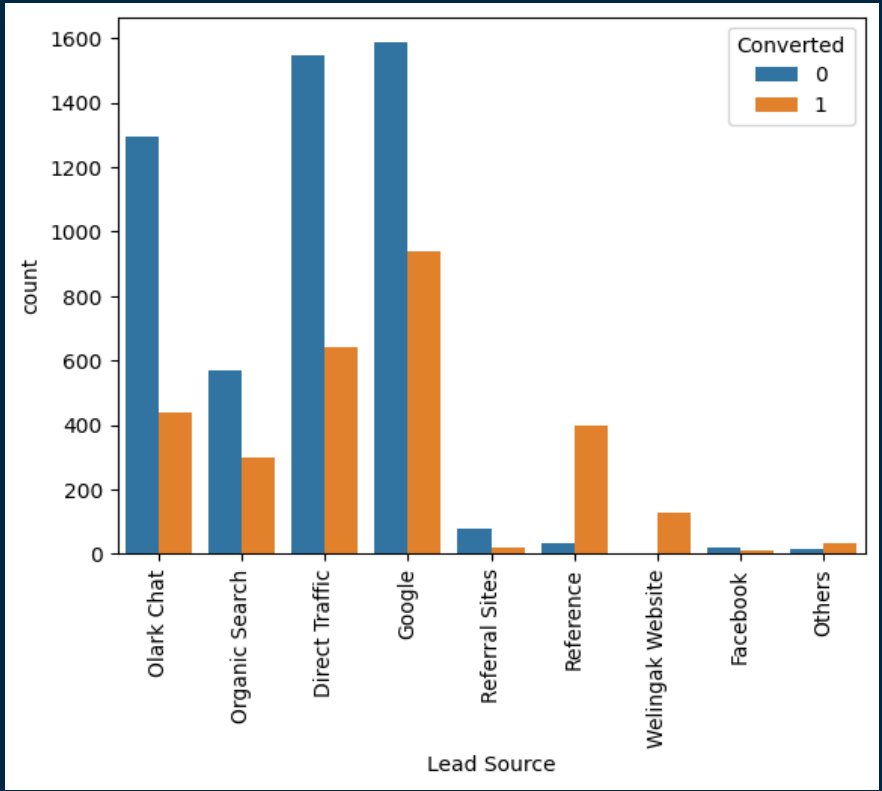


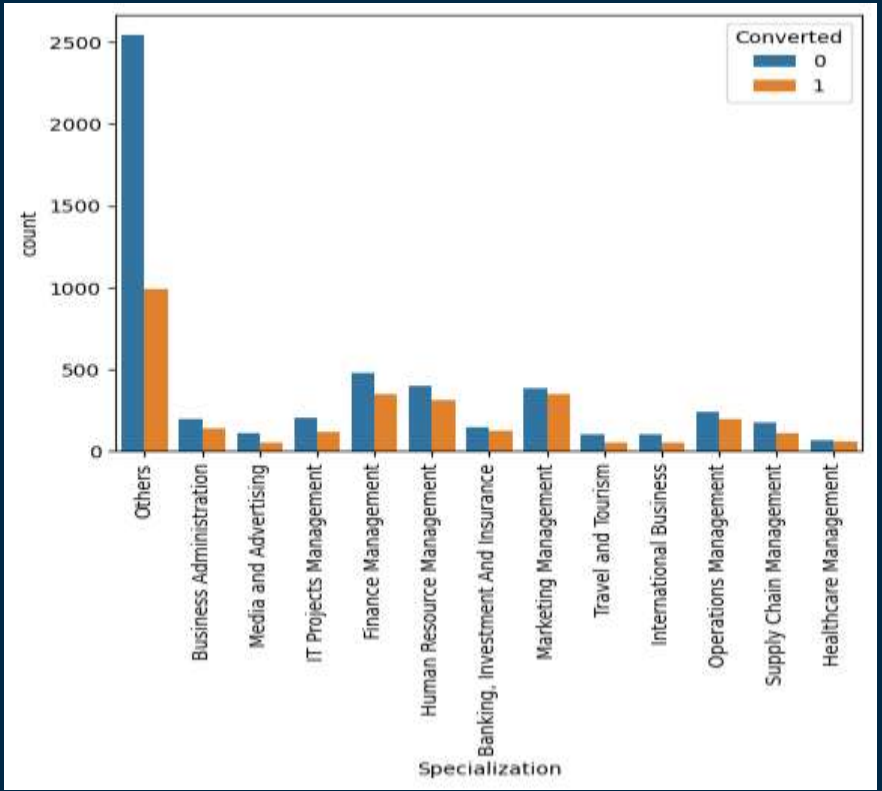
Bivariate Analysis



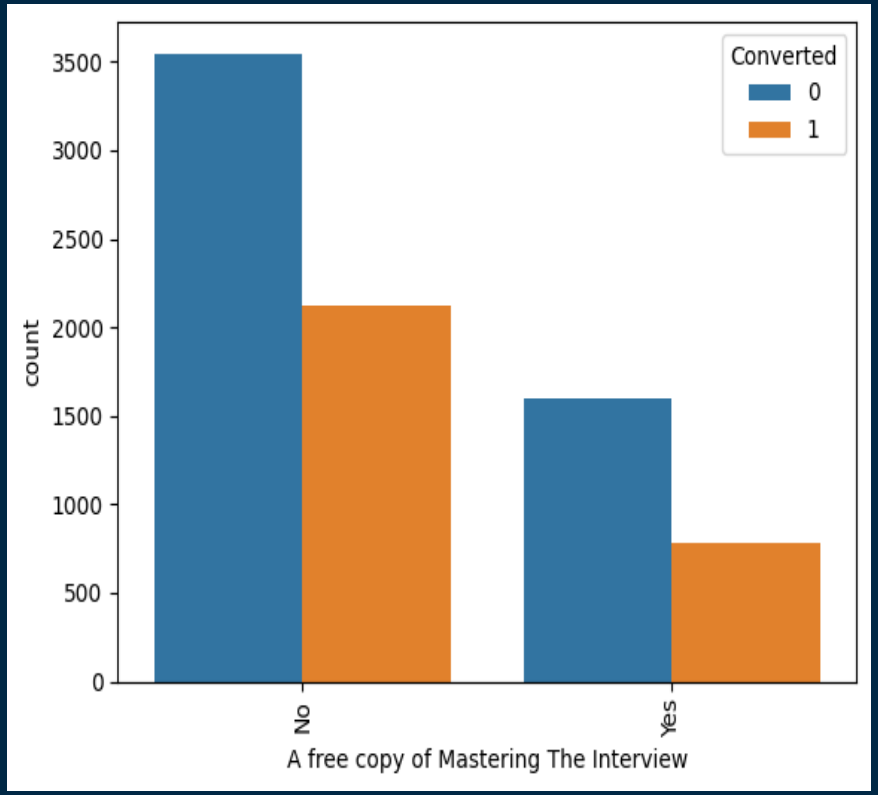


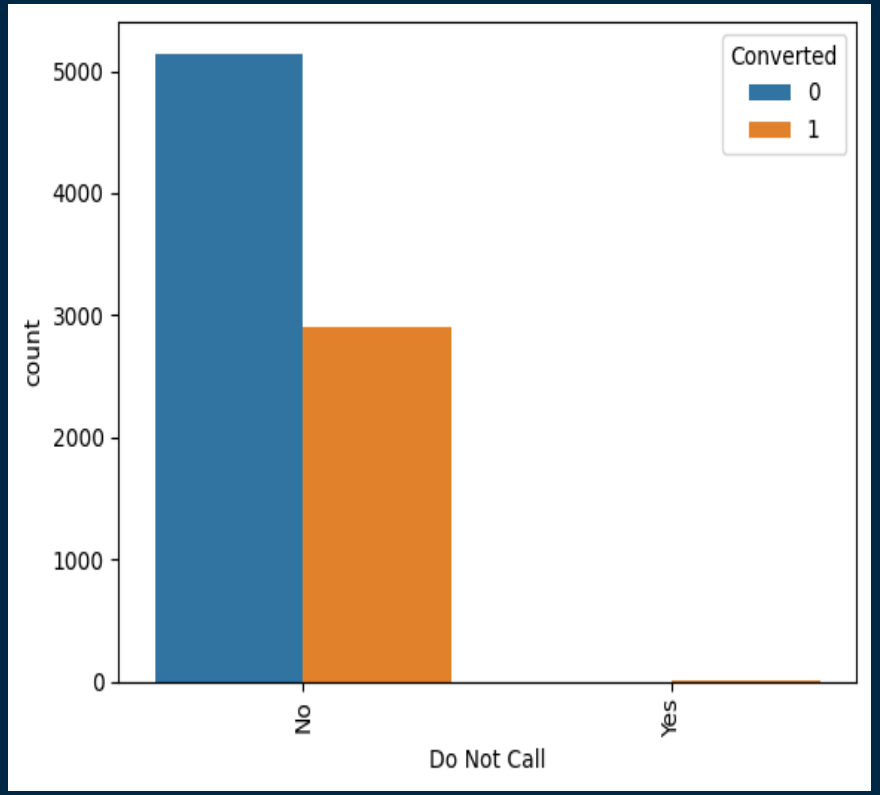
Bivariate Analysis





Bivariate Analysis

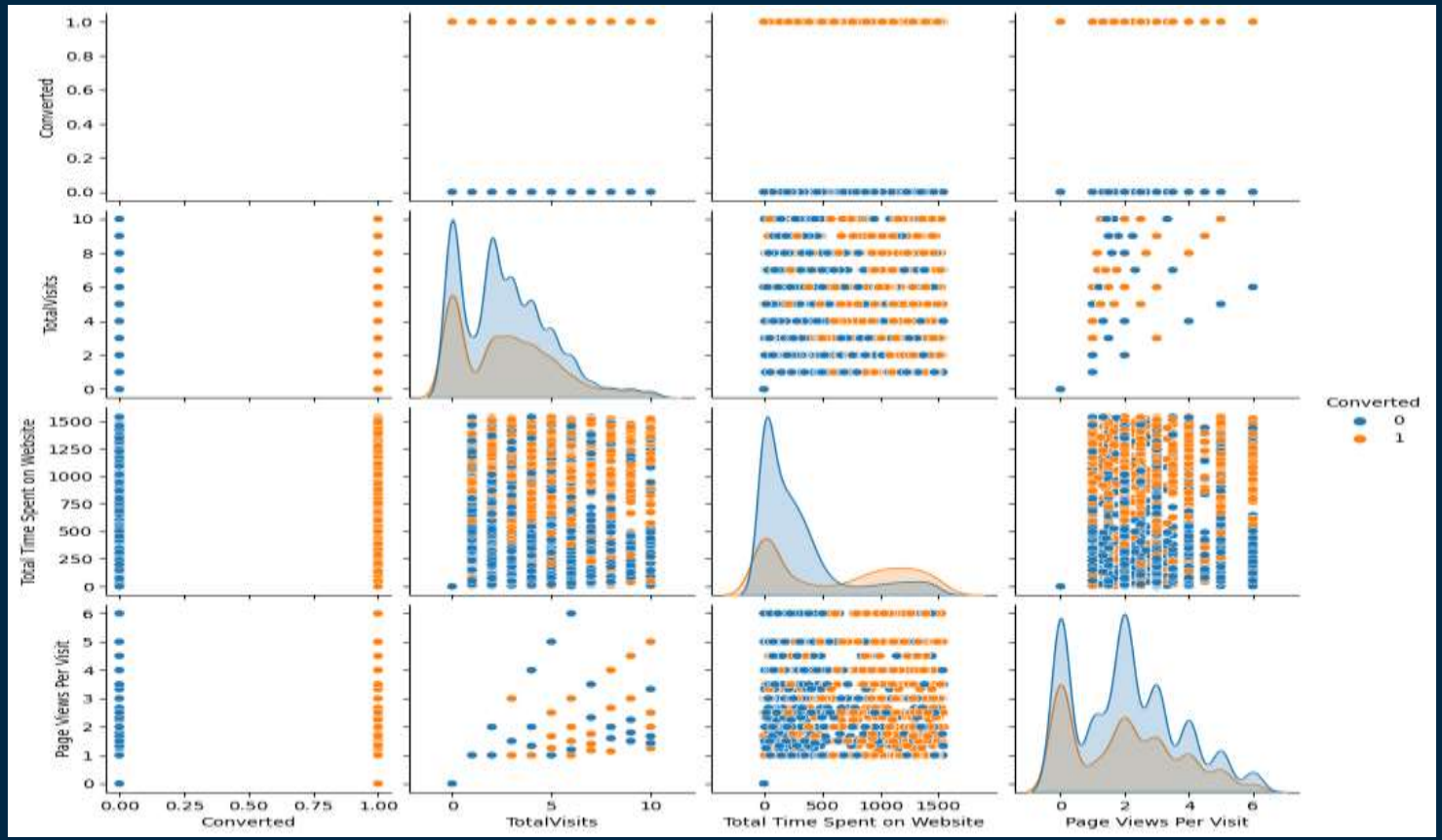




Bivariate Analysis

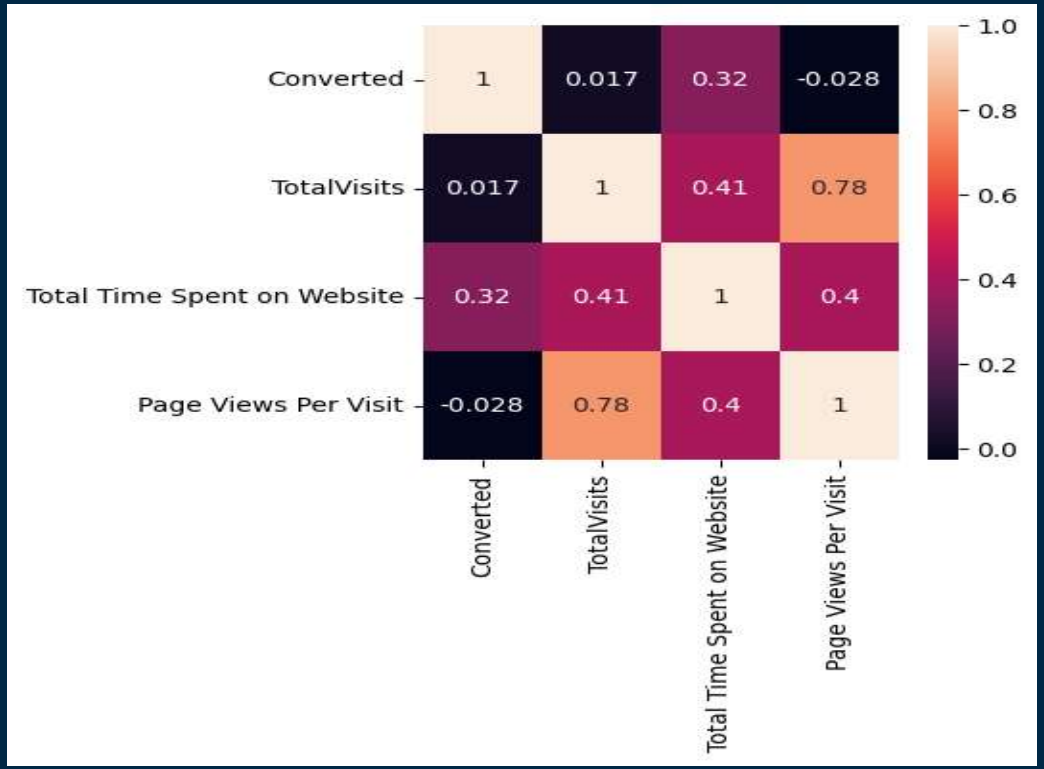
- Leads originated from Landing page submission and Lead add Forms have most conversions.
- Google & Direct Traffic are the top most Lead Source of conversions.
- Customers whose last activity was either Email opened or SMS Sent have highest conversions
- Do Not Email column does not provide much information about lead conversion.
- India has the major portion of the conversions with respect to other countries.
- Finance, marketing & HR mgmt are the specialization after others to be focussed for conversions.
- Working professionals should be targeted in priority as they have the highest ratio of conversion while Housewives are the lowest.
- Better career prospects is the main reason for opting a course & clients with this reason also can be given preference.
- Tags of 'revert after reading the email' has the highest lead conversion rate amoung other tags.
- Mumbai city having highest conversion rate should be targeted first for lead conversion at priority while Tier II cities the least.
- SMS sent has the highest conversion rate followed by Email opened as their last notable activity.

EDA



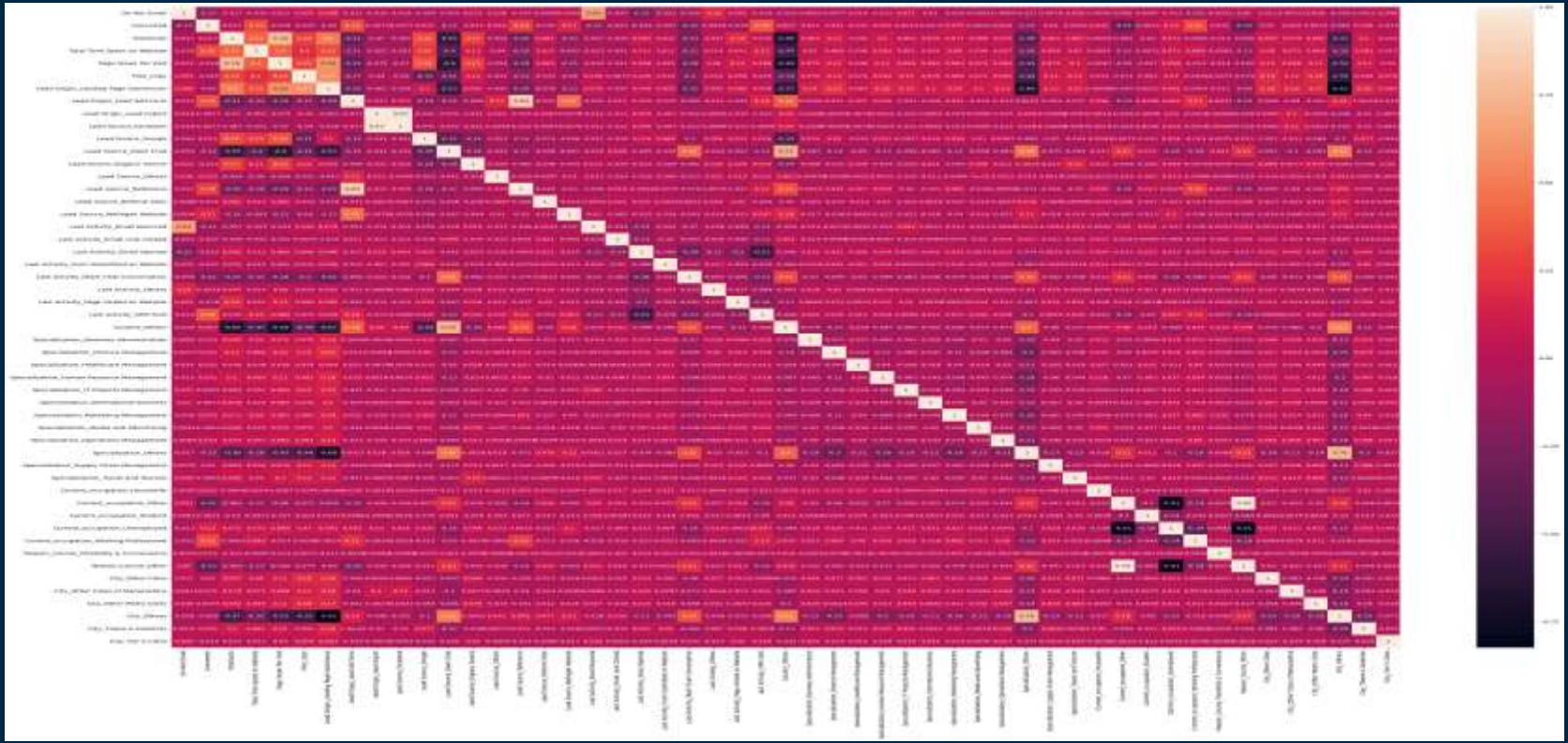
Bivariate Analysis for Numerical Variables

HeatMap



correlation between numerical variables before dummy variable creation and standardization

Correlation Matrix for numerical variables



MODEL BUILDING

- Splitting into train and test set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high p-values
- Check VIF value for all the existing columns
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test predictions

Model 1

Model 1

```
import statsmodels.api as sm
# Adding a constant variable from X_train dataframe with variables selected by RFE
X_train_sm=sm.add_constant(X_train[cols])
# Create a fitted model
logreg_model1 = sm.GLM(y_train, X_train_sm,family=sm.families.Binomial())
res = logreg_model1.fit()
res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	5635
Model:	GLM	**Df Residuals:**	5619
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2114.0
Date:	Sat, 13 Jan 2024	**Deviance:**	4227.9
Time:	17:09:35	**Pearson chi2:**	5.63e+03
No. Iterations:	23	**Pseudo R-squ. (CS):**	0.4229
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5805	199.894	-0.003	0.998	-392.366	391.205
Do Not Email	-0.3790	0.055	-6.910	0.000	-0.486	-0.271
TotalVisits	0.3696	0.060	6.144	0.000	0.252	0.487
Total Time Spent on Website	1.0611	0.044	24.159	0.000	0.975	1.147
Page Views Per Visit	-0.3097	0.068	-4.563	0.000	-0.443	-0.177
Lead Origin_Landing Page Submission	-0.4301	0.064	-6.695	0.000	-0.556	-0.304
Lead Origin_Lead Add Form	0.8658	0.066	13.124	0.000	0.736	0.995
Lead Source_Olark Chat	0.5470	0.064	8.521	0.000	0.421	0.673
Lead Source_Welingak Website	2.7946	1605.680	0.002	0.999	-3144.279	3149.869
Last Activity_Email Opened	0.3644	0.061	6.016	0.000	0.246	0.483
Last Activity_Olark Chat Conversation	-0.2525	0.068	-3.720	0.000	-0.386	-0.119
Last Activity_Others	0.2183	0.039	5.669	0.000	0.143	0.294
Last Activity_SMS Sent	0.9748	0.059	16.643	0.000	0.860	1.090
Specialization_Others	-0.3270	0.057	-5.738	0.000	-0.439	-0.215
Current_occupation_Other	-0.5931	0.046	-12.994	0.000	-0.683	-0.504
Current_occupation_Working Professional	0.7113	0.056	12.724	0.000	0.602	0.821

Model 2

Model 2

In [81]:
```
import statsmodels.api as sm
# Adding a constant variable from X_train dataframe with variables selected by RFE
X_train_sm=sm.add_constant(X_train[cols])
# Create a fitted model
logreg_model1 = sm.GLM(y_train, X_train_sm,family=sm.families.Binomial())
res = logreg_model1.fit()
res.summary()
```

Out[81]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5635
Model:	GLM	Df Residuals:	5620
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2161.4
Date:	Fri, 12 Jan 2024	Deviance:	4322.7
Time:	20:19:23	Pearson chi2:	5.93e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4131
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1862	0.160	-7.396	0.000	-1.501	-0.872
Total Time Spent on Website	1.0730	0.044	24.631	0.000	0.988	1.158
Lead Origin_Landing Page Submission	-0.8933	0.127	-7.041	0.000	-1.142	-0.645
Lead Origin_Lead Add Form	3.5913	0.228	15.754	0.000	3.145	4.038
Lead Source_Olark Chat	1.2686	0.131	9.654	0.000	1.011	1.526
Last Activity_Email Bounced	-1.6845	0.449	-3.748	0.000	-2.565	-0.804
Last Activity_Email Opened	0.7349	0.125	5.871	0.000	0.490	0.980
Last Activity_Olark Chat Conversation	-0.8063	0.211	-3.826	0.000	-1.219	-0.393
Last Activity_Others	1.0176	0.263	3.865	0.000	0.502	1.534
Last Activity_SMS Sent	1.9835	0.128	15.517	0.000	1.733	2.234
Specialization_Media and Advertising	-0.6496	0.260	-2.496	0.013	-1.160	-0.140
Specialization_Others	-0.6839	0.114	-6.011	0.000	-0.907	-0.461
Current_occupation_Other	-1.2695	0.098	-12.966	0.000	-1.461	-1.078
Current_occupation_Working Professional	2.6867	0.211	12.738	0.000	2.273	3.100
City_Tier II Cities	-0.6206	0.456	-1.361	0.174	-1.514	0.273

Key Take away:

'City_Tier II Cities' has high p-value of 0.174.Therefore this column will be removed from the model.

Model 3

Model 3

In [83]:
```
import statsmodels.api as sm
# Adding a constant variable from X_train dataframe with variables selected by RFE
X_train_sm=sm.add_constant(X_train[cols])
# Create a fitted model
logreg_model1 = sm.GLM(y_train, X_train_sm,family=sm.families.Binomial())
res = logreg_model1.fit()
res.summary()
```

Out[83]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	5635
Model:	GLM	Df Residuals:	5621
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2162.3
Date:	Fri, 12 Jan 2024	Deviance:	4324.6
Time:	20:19:41	Pearson chi2:	5.90e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4129
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1868	0.160	-7.401	0.000	-1.501	-0.872
Total Time Spent on Website	1.0721	0.044	24.625	0.000	0.987	1.157
Lead Origin_Landing Page Submission	-0.9025	0.127	-7.127	0.000	-1.151	-0.654
Lead Origin_Lead Add Form	3.5897	0.228	15.749	0.000	3.143	4.036
Lead Source_Olark Chat	1.2672	0.131	9.646	0.000	1.010	1.525
Last Activity_Email Bounced	-1.6860	0.449	-3.754	0.000	-2.566	-0.806
Last Activity_Email Opened	0.7381	0.125	5.896	0.000	0.493	0.983
Last Activity_Olark Chat Conversation	-0.8044	0.211	-3.817	0.000	-1.217	-0.391
Last Activity_Others	1.0243	0.263	3.892	0.000	0.508	1.540
Last Activity_SMS Sent	1.9824	0.128	15.508	0.000	1.732	2.233
Specialization_Media and Advertising	-0.6419	0.260	-2.467	0.014	-1.152	-0.132
Specialization_Others	-0.6848	0.114	-6.026	0.000	-0.908	-0.462
Current_occupation_Other	-1.2679	0.098	-12.949	0.000	-1.460	-1.076
Current_occupation_Working Professional	2.6893	0.211	12.745	0.000	2.276	3.103

Key Take away:

'Specialization_Media and Advertising' has p-value of 0.014.Therefore this column will be removed from the model.

Model 4

Model 4

In [84]:
```
import statsmodels.api as sm
# Adding a constant variable from X_train dataframe with variables selected by RFE
X_train_sm=sm.add_constant(X_train[cols])
# Create a fitted model
logreg_model1 = sm.GLM(y_train, X_train_sm,family=sm.families.Binomial())
res = logreg_model1.fit()
res.summary()
```

Out[84]:

Generalized Linear Model Regression Results

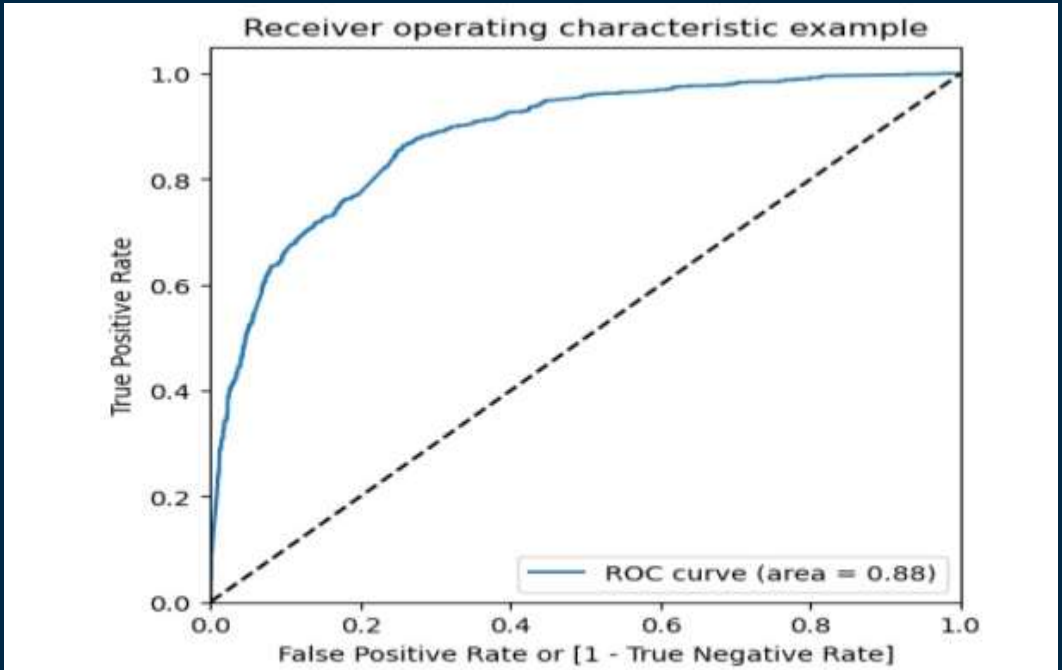
Dep. Variable:	Converted	No. Observations:	5635
Model:	GLM	Df Residuals:	5622
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2165.5
Date:	Fri, 12 Jan 2024	Deviance:	4331.0
Time:	15:27:47	Pearson chi2:	5.88e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4122
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2106	0.160	-7.569	0.000	-1.524	-0.897
Total Time Spent on Website	1.0722	0.043	24.658	0.000	0.987	1.157
Lead Origin_Landing Page Submission	-0.9018	0.126	-7.130	0.000	-1.150	-0.654
Lead Origin_Lead Add Form	3.5950	0.228	15.782	0.000	3.149	4.041
Lead Source_Olark Chat	1.2638	0.131	9.634	0.000	1.007	1.521
Last Activity_Email Bounced	-1.6768	0.448	-3.743	0.000	-2.555	-0.799
Last Activity_Email Opened	0.7395	0.125	5.908	0.000	0.494	0.985
Last Activity_Olark Chat Conversation	-0.8058	0.211	-3.823	0.000	-1.219	-0.393
Last Activity_Others	1.0154	0.262	3.871	0.000	0.501	1.530
Last Activity_SMS Sent	1.9740	0.128	15.453	0.000	1.724	2.224
Specialization_Others	-0.6578	0.113	-5.821	0.000	-0.879	-0.436
Current_occupation_Other	-1.2652	0.098	-12.938	0.000	-1.457	-1.074
Current_occupation_Working Professional	2.6754	0.210	12.712	0.000	2.263	3.088

Model 4 seems stable and has acceptable p-values within the threshold which can be used for further analysis.

Model Evaluation

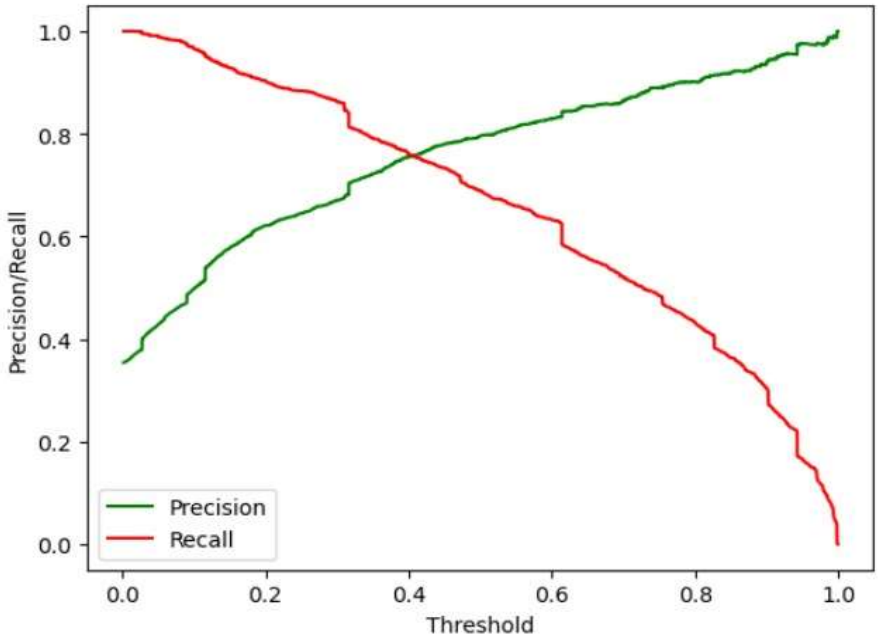
ROC curve for test dataset



Key Takeaways:

Area under the curve for test dataset is 0.88.

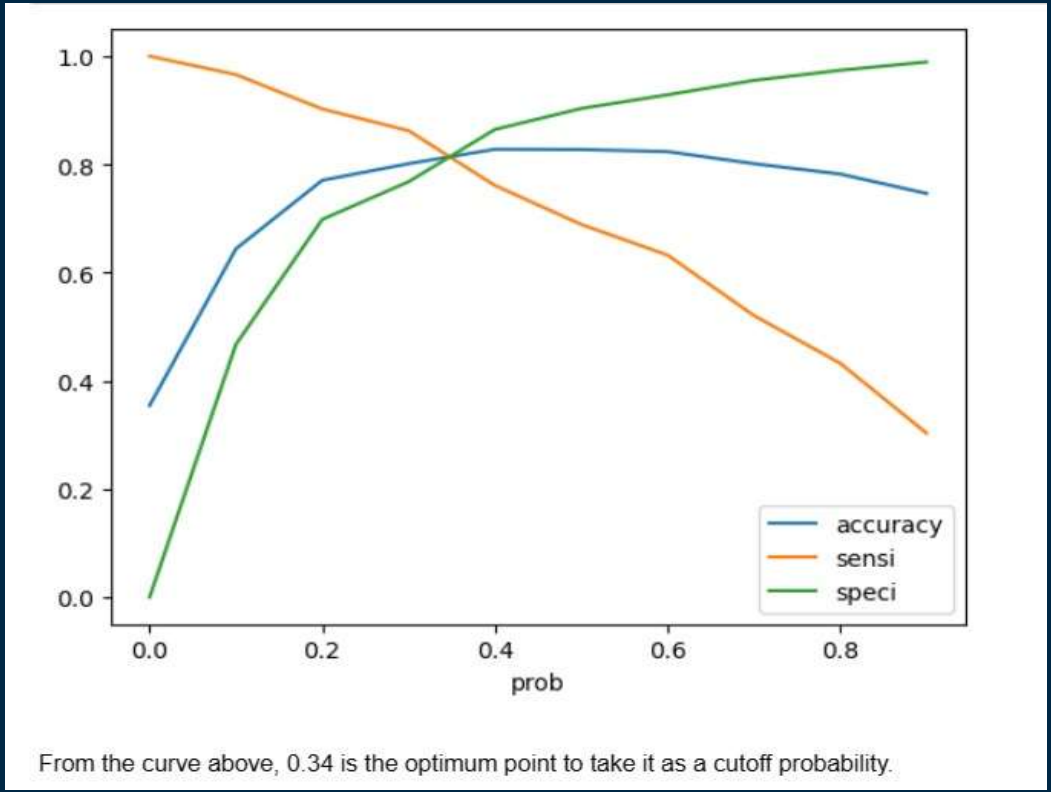
Precision and recall tradeoff for Train dataset



Key Takeaways:

The precision-recall tradeoff value is approximately 0.4 with a balance between precision and recall.

Accuracy, Sensitivity and Specificity



From the curve above, 0.34 is the optimum point to take it as a cutoff probability.

Final features selected through RFE with their respective coefficients

Lead Origin_Lead Add Form	3.595028
Current_occupation_Working Professional	2.675383
Last Activity_SMS Sent	1.974016
Lead Source_Olark Chat	1.263781
Total Time Spent on Website	1.072192
Last Activity_Others	1.015430
Last Activity_Email Opened	0.739538
Specialization_Others	-0.657774
Last Activity_Olark Chat Conversation	-0.805813
Lead Origin_Landing Page Submission	-0.901833
const	-1.210571
Current_occupation_Other	-1.265202
Last Activity_Email Bounced	-1.676832

dtype: float64

Final Conclusion

Evaluation Metrics

Train set :

- Accuracy -> 81.7%
- Sensitivity-> 79.9%
- Specificity-> 82.7%

For Test set :

- Accuracy : 79.8%
- Sensitivity : 75.99%
- Specificity : 82.15%

Evaluation metrics in both test and train dataset are consistent. Therefore final model is performing good.

Top 3 features contributing to predicting hot leads are:

- Lead Origin_Lead Add Form
- Current_occupation_Working Professional
- Last Activity_SMS Sent

Recommendations

1. Top three variables in model which contribute most towards the probability of a lead getting converted are:
- Lead Origin
- Current_occupation
- Last Activity

2. Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion are :
- Lead Origin of Lead Add Form
- Working Professional as occupation
- If the last activity by Customer was SMS sending

3. Areas of improvement are Specialization-Others ,Last Activity of Olark Chat conversation Last Activity where email bounced

Thank You