LEAD SCORING CASE STUDY

Presented by:

- Ashish Arora
- Hitesh S Jadhav
- Supreetha R

Problem Statement

- X Education is an education company that sells online courses to industry professionals.
- X Education wants to select most promising leads that can be converted to paying customers.
- Although the company generates a lot of leads only a few are converted into paying customers, wherein the company wants a higher lead conversion. Leads come through numerous modes like email, advertisements on websites, google searches etc.
- The company has had 30% conversion rate through the whole process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. The implementation process of lead generating attributes are not efficient in helping conversions.

Business Goals

- The company requires a model to be built for selecting most promising leads. Lead score to be given to each leads such that it indicates how promising the lead could be.
- The higher the lead score the more promising the lead to get converted, the lower it is the lesser the chances of conversion.
- The model to be built should have lead conversion rate around 80% or more.

Work Strategy

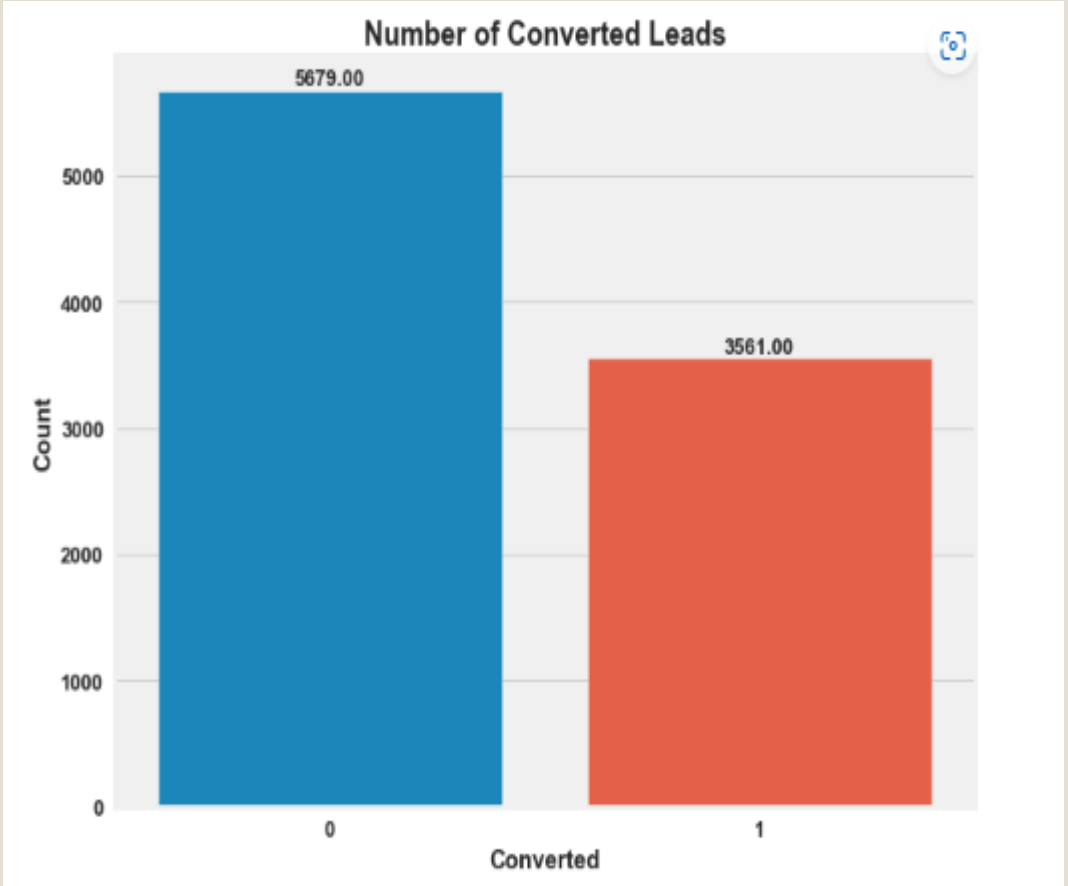
- ❖Import Libraries
- ❖Data Loading and Reading
- ❖Exploratory Data Analysis
- ❖Data Preparation For Modelling
- ❖Model Building
- ❖Model Evaluation: Train
- ❖Model Evaluation: Test
- ❖Insights and Recommendations

Importing Libraries

NumPy	For Mathematical Functions
pandas	For Data Manipulation and Analysis
matplotlib	For Visualizations
seaborn	For Visualizations
statsmodels	For Estimation of Statistical Model
scikit learn	For Predictive Analysis

Data Understanding

- ❖The data frame contains
  - ✓36 Independent variable.
  - ✓1 target variable.
  - ✓9240 rows.
- ❖The data types of features are as follows:
  - ✓4 features are float64.
  - ✓3 features are integer.
  - ✓30 features are object.
- ❖There are lots of features with null values.
- ❖Some data values also have select values; those will also be considered as null.
- ❖Target data is Imbalanced.

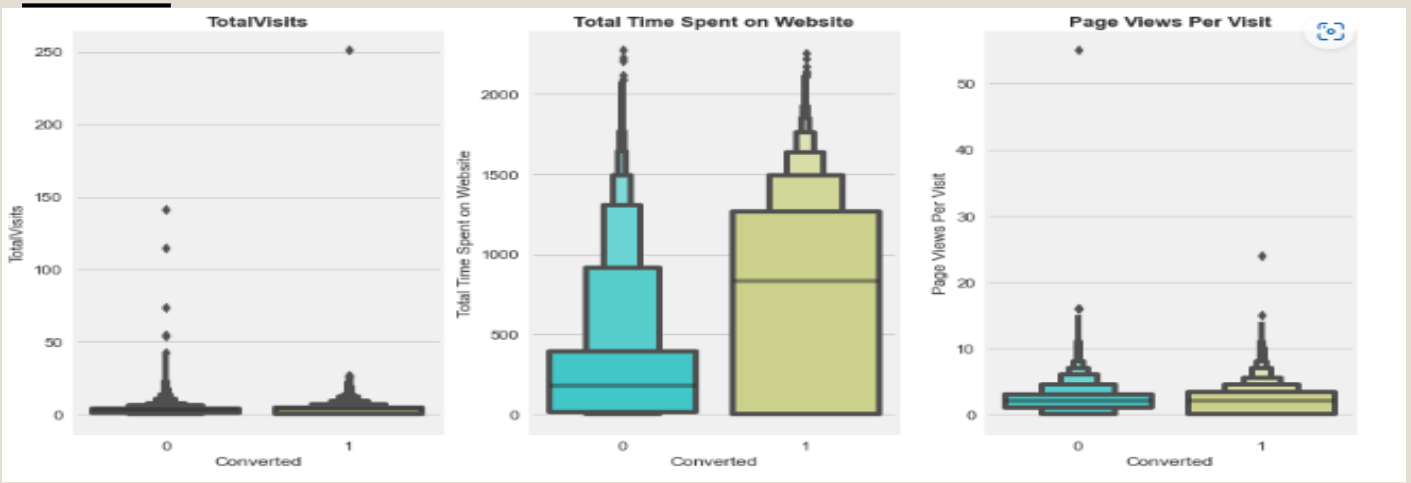


Exploratory Data Analysis

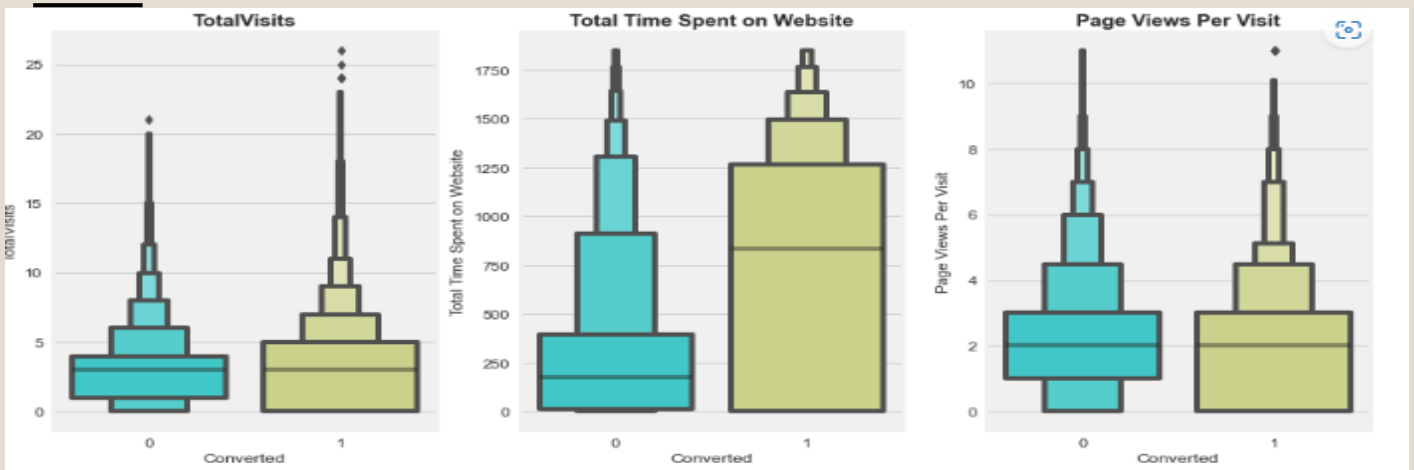
Univariate Analysis: Data Cleaning

- **Treated Missing Values**
- **Treated Outliers**
  - ✓Capping
  - ✓Trimming
- **Combined various columns values having same meaning.**
  - ✓Tags
  - ✓Specialization
  - ✓Last Activity
  - ✓Lead Source
  - ✓Country
  - ✓City
  - ✓What is Your current occupation

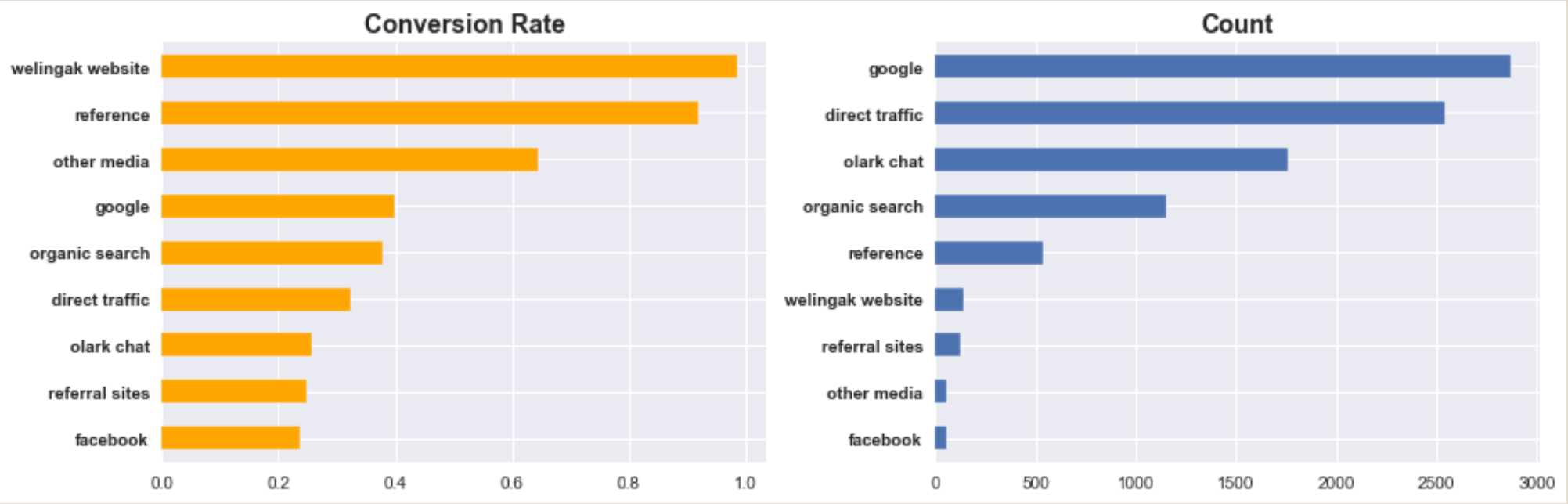
Before



After



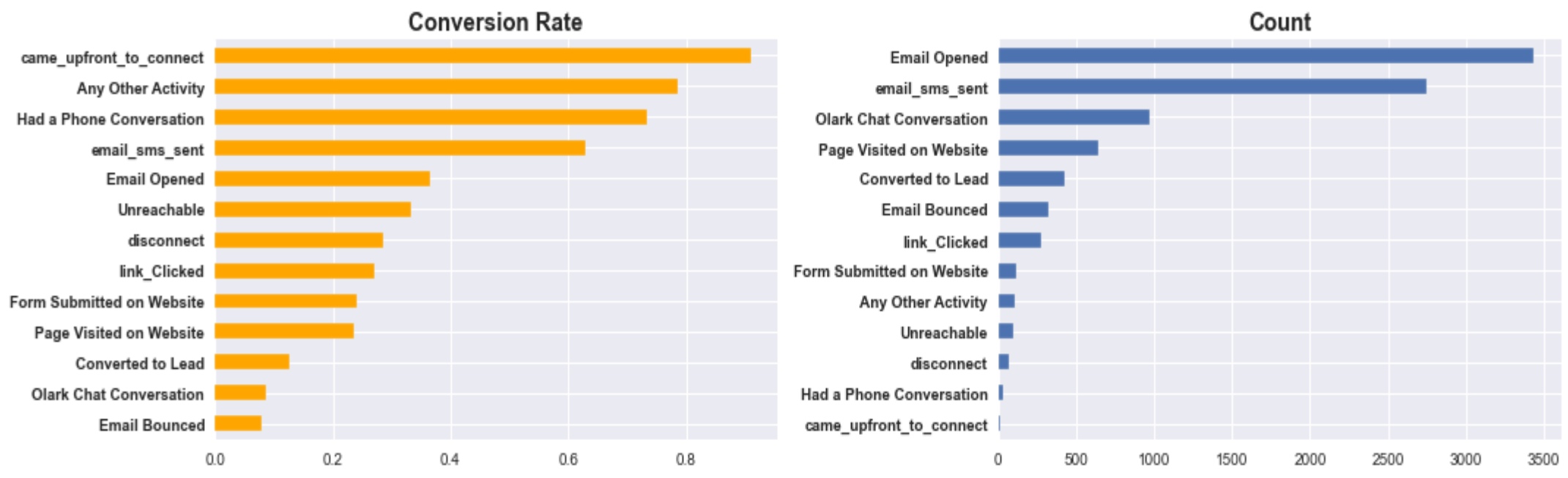
Bivariate Analysis– Lead Source



Insights:

- Welingak website and reference contribute most to the conversion rate, but count from these sources are less in number.
- Google is the most prominent source of getting leads but the conversion rate of it is only 40%

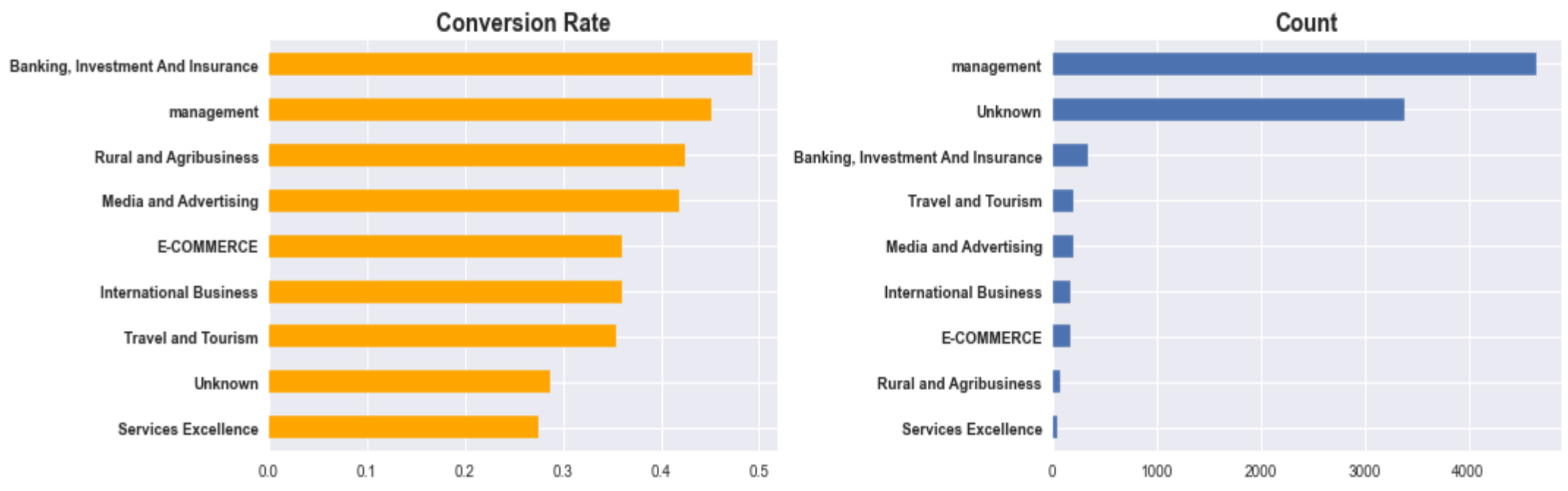
Bivariate Analysis – Last Activity



INSIGHTS:

- Clients who come upfront has better chance of becoming a hot lead though such clients are less in number. Same as the case for Had a phone conversation.
- Clients who submitted the form, tags such as converted to lead, Olark chat conversation have poor conversion rate.

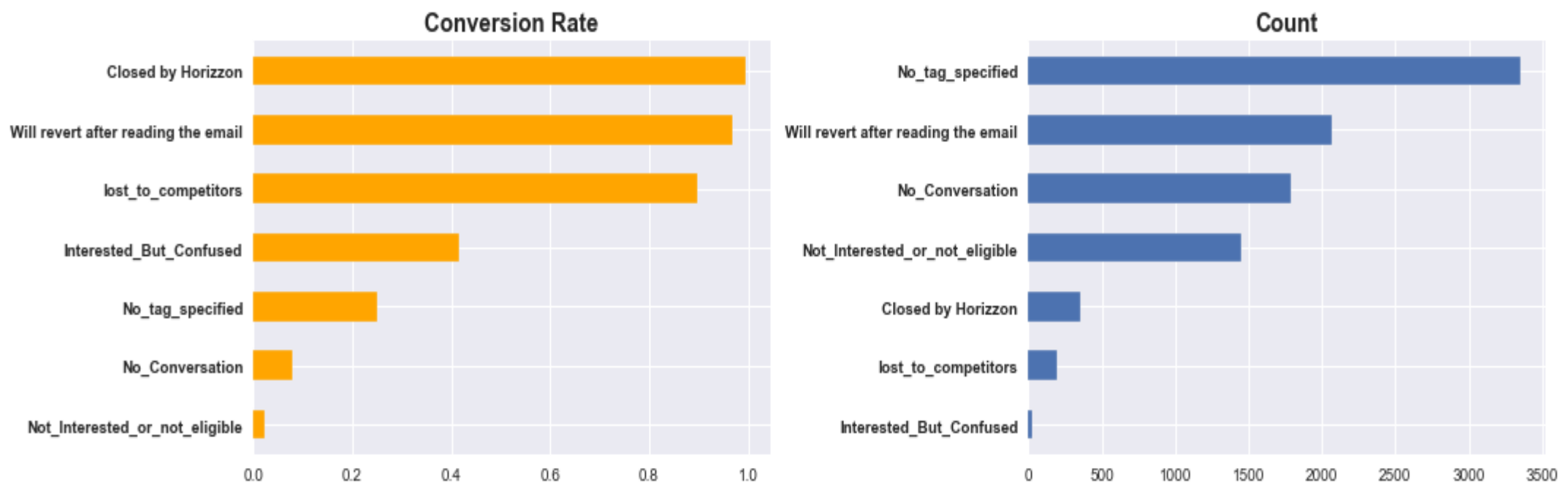
Bivariate Analysis – Specialization



INSIGHTS

- Clients which are in Service excellence has the least interest towards the given course as they are less in numbers as well as their response rate is poor than the others.
- Clients who are in management are most favorable people to opt for this course. They are good in leads as well as response rate.
- clients who are from Banking, Investment and Insurance Domain are the most prominent leads.

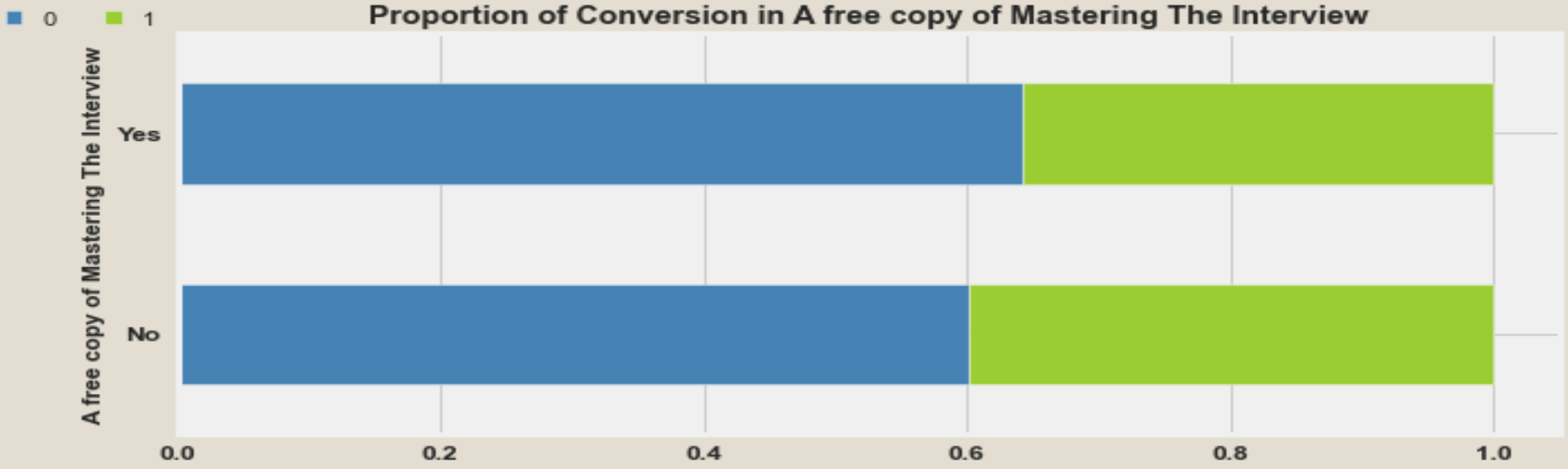
Bivariate Analysis – Tags



INSIGHTS

- Clients which are closed by Horizons or lost to competitors have the highest conversion rate, but these clients are very limited.
- Clients who reverts to email are also prominent.
- Interested but confused can be the become a lead.

Bivariate Analysis: A free copy of Mastering the Interview



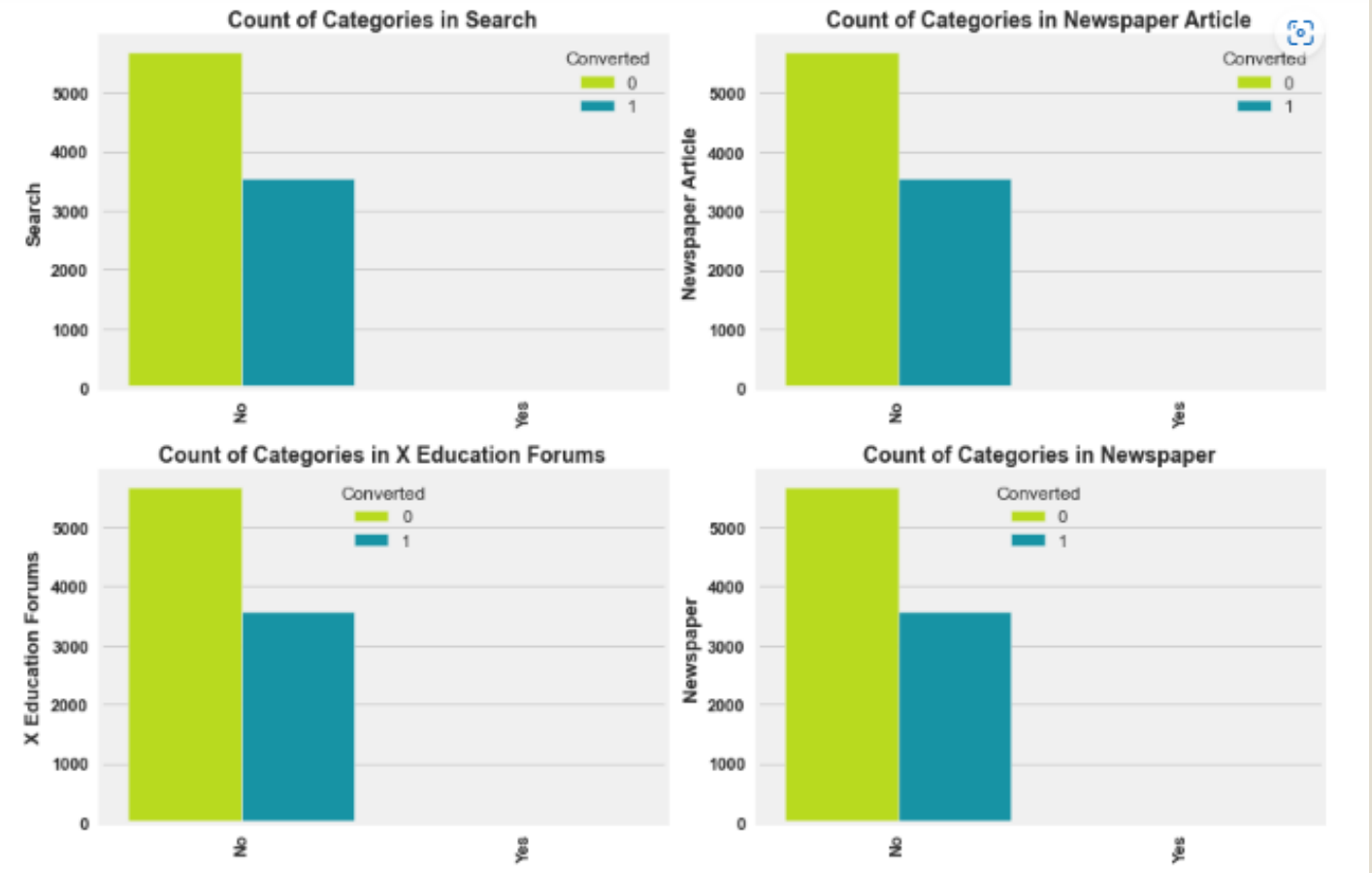
INSIGHTS:

- A free copy of Mastering the Interview is not giving edge to the organization for converting to the lead.
- It is suggested for the organization to not give away this. It would just increase the cost.
- It would be better to give the leads who are more prone to conversion as added benefits rather than freebie.

EDA– Dropping Features

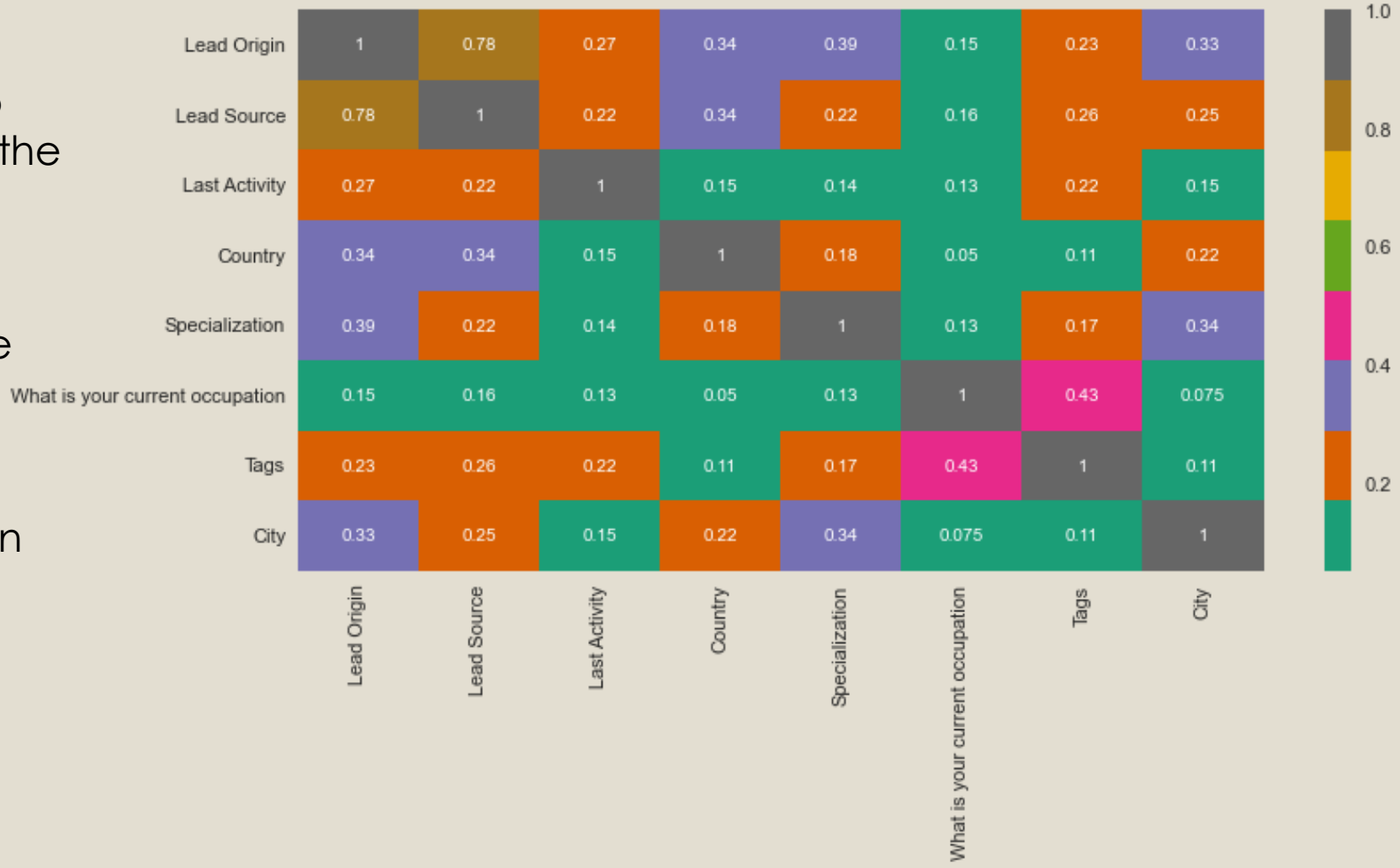
Following columns doesn't add any values to the study as they signifies no information to the study. Hence Dropped them.

- ❖'Do Not Email'
- ❖'Do Not Call'
- ❖'Search'
- ❖'Newspaper Article'
- ❖'X Education Forums'
- ❖'Newspaper'
- ❖'Digital Advertisement'
- ❖'Through Recommendations'



Multivariate Analysis: Correlation within Nominal Varaible

- Using CramersV, We tried to find the correlation among the categorical Variables
- Only lead origin and lead source seems to have more relation with each other of 0.78.
- Rest other shares less relation with each other



Multivariate Analysis: Continuous Variables

- Total visits and Page per visits shows a positive trend.
- The trend seems to have multiple situations and clusters.
- Clients which were converted have spent more time on the websites especially around 1200 to 2000
- For some clients Page Views per Visit doesn't mattered with the amount of total time spent on websites.

Modeling

Data preparation:

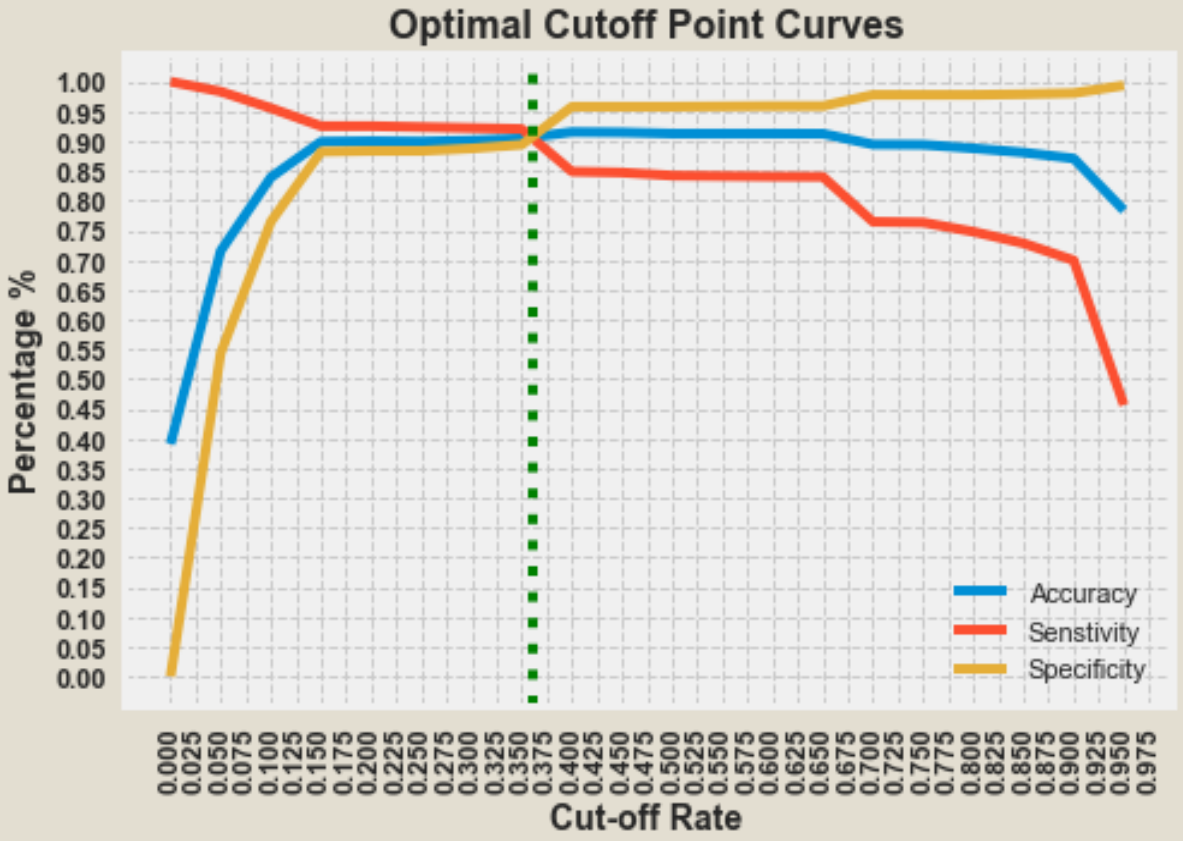
- ❖ **Dummy Variables**
  - ✓ **There were total of 9 categorical columns.**
  - ✓ **57 dummy variables are formed out of them.**
- ❖ **Train Test Split**
  - ✓ **We split the data in 70-30 ratio of train-test respectively.**
- ❖ **Standard Scaling**
  - ✓ **We used standard scaling for model**
- ❖ **Dropped Dummy variables whose information were not precisely given.**
- ❖ **Variance inflation factor**
  - ✓ **During EDA, We realized there were some columns which have collinearity among them. To check it we used VIF and there exists some multicollinearity.**

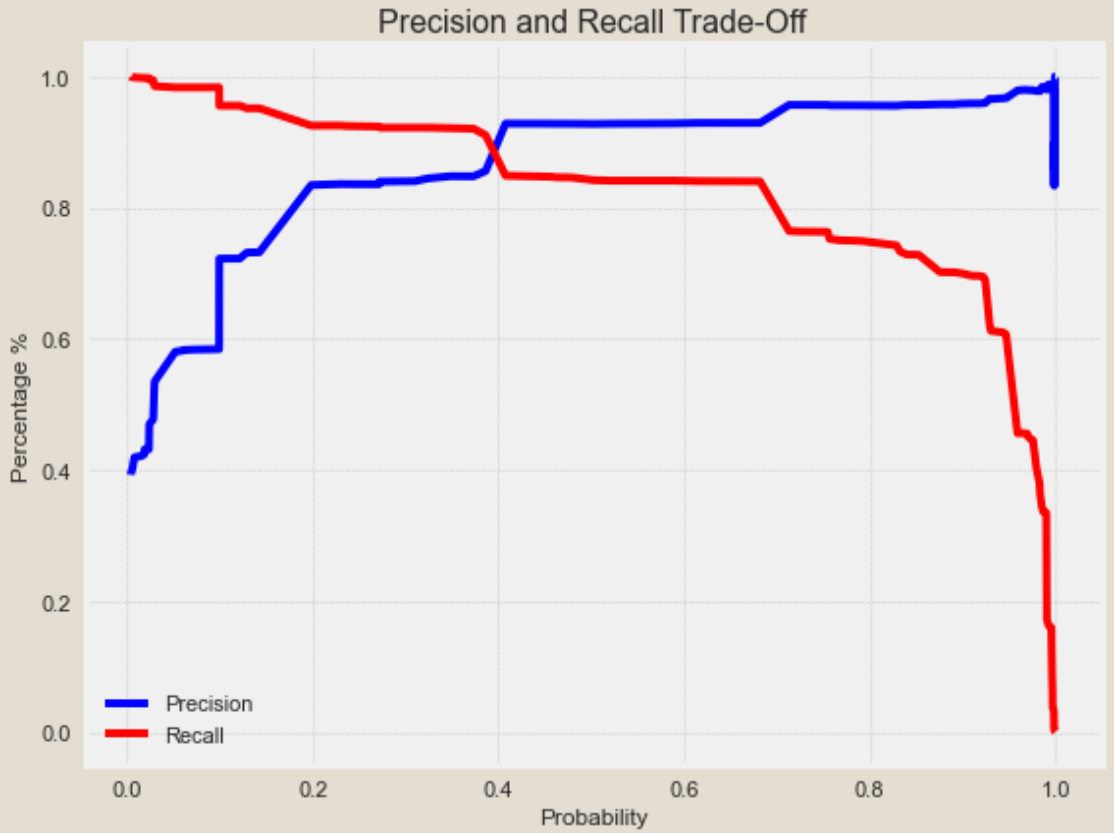
Model Building

- Using Recursive Feature Elimination, we selected 15 features.
- Using these 15 features, we started building our models.
- While building models, we eliminated 3 features based on high p-values.
- Our final model has 12 features.
- All features in the model have VIF values of less than 2.
- All 12 features of our final model are significant.
- These are the statistics of our final model.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2048	0.093	-23.795	0.000	-2.386	-2.023
Lead Origin_Lead Add Form	0.9946	0.235	4.227	0.000	0.533	1.456
Lead Source_welingak website	2.3037	1.038	2.219	0.026	0.269	4.338
Last Activity_Converted to Lead	-1.4973	0.322	-4.655	0.000	-2.128	-0.867
Last Activity_Email Bounced	-1.7494	0.331	-5.292	0.000	-2.397	-1.102
Last Activity_Olark Chat Conversation	-1.2808	0.200	-6.410	0.000	-1.672	-0.889
Last Activity_email_sms_sent	1.7424	0.113	15.459	0.000	1.521	1.963
What is your current occupation_Student	2.5581	0.445	5.745	0.000	1.685	3.431
What is your current occupation_Unemployed	2.9678	0.127	23.324	0.000	2.718	3.217
What is your current occupation_Working Professional	3.7841	0.299	12.667	0.000	3.199	4.370
Tags_No_Conversation	-4.2989	0.164	-26.172	0.000	-4.621	-3.977
Tags_Not_Interested_or_not_eligible	-4.4834	0.237	-18.921	0.000	-4.948	-4.019
Tags_Will revert after reading the email	2.1214	0.181	11.696	0.000	1.766	2.477
Tags_lost_to_competitors	3.3291	0.349	9.535	0.000	2.645	4.013

Final Model Evaluation: Optimal Cut-Off

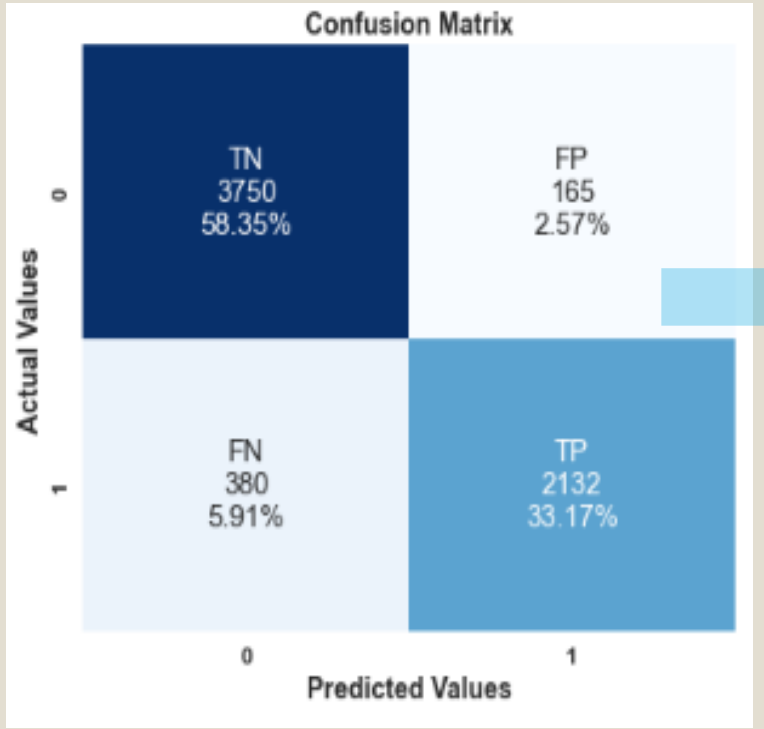




- **After considering, both the curves we come up with a optimal cut off rate of 0.39**

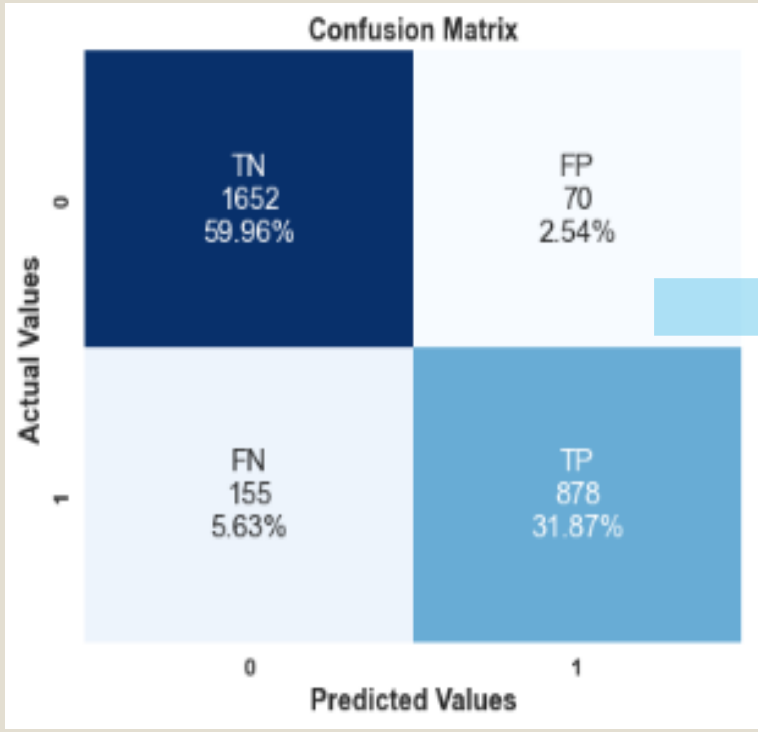
Model Evaluation: Confusion Matrix

Final Model: Train Scores



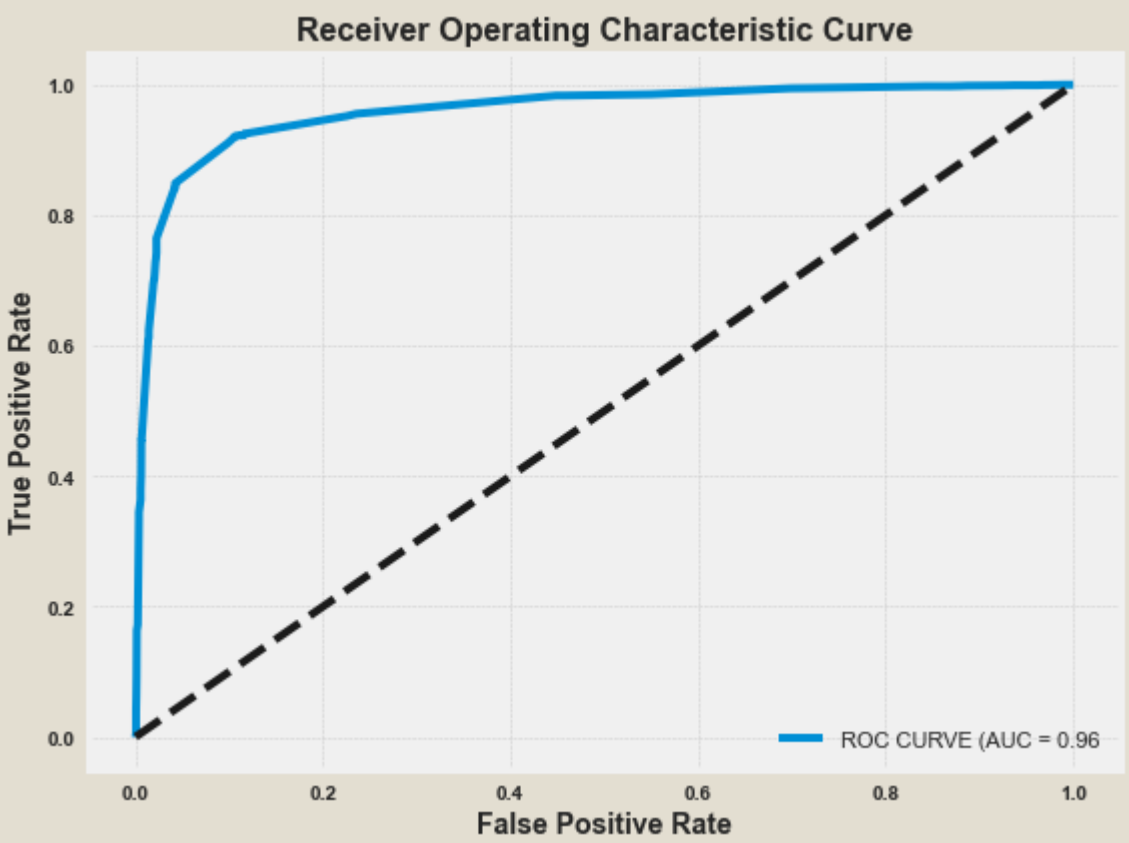
	train_mod_3
Accuracy	91.52
Senstivity	84.87
Specificity	95.78
FPR	4.21
FNR	15.13
Recall	84.87
Precision	92.82
F1_score	88.67
AUC	95.94
Num of Features	13.00

Final Model: Test Scores

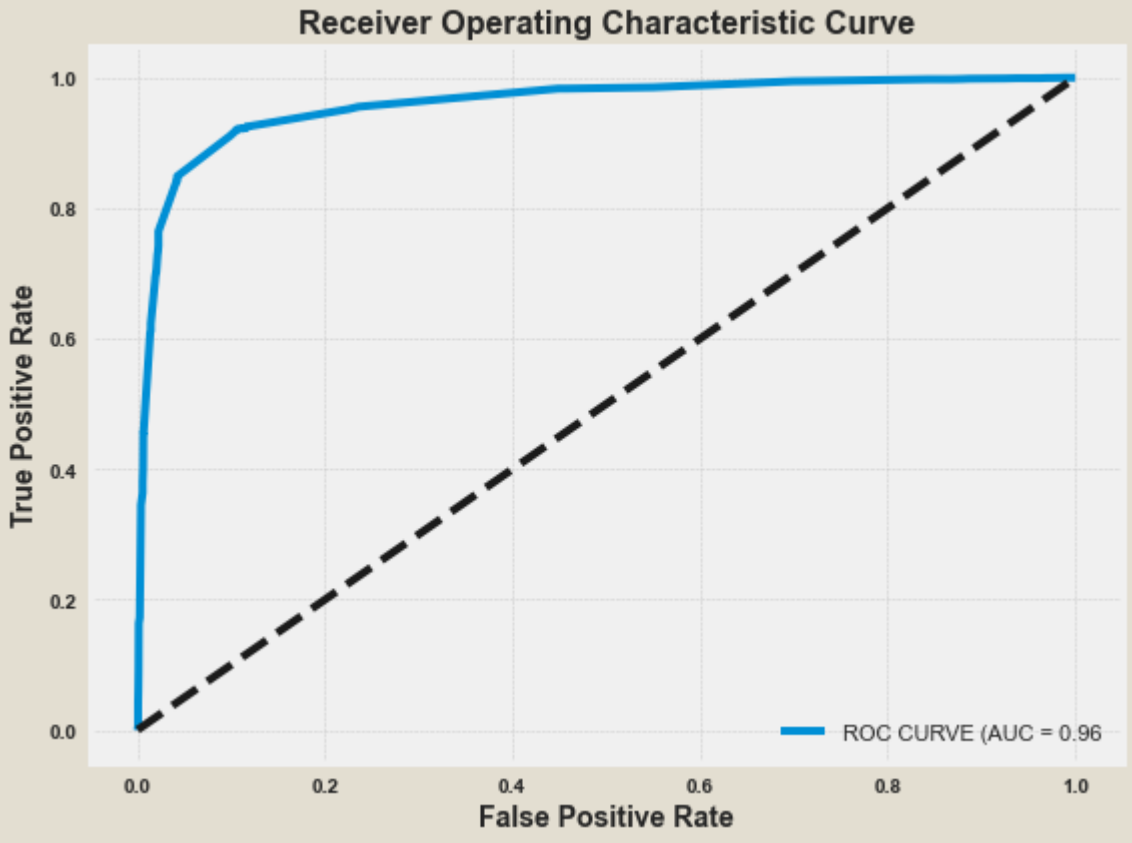


	test_mod_3
Accuracy	91.83
Senstivity	85.00
Specificity	95.94
FPR	4.07
FNR	15.01
Recall	85.00
Precision	92.62
F1_score	88.64
AUC	95.60
Num of Features	13.00

Model Evaluation: ROC

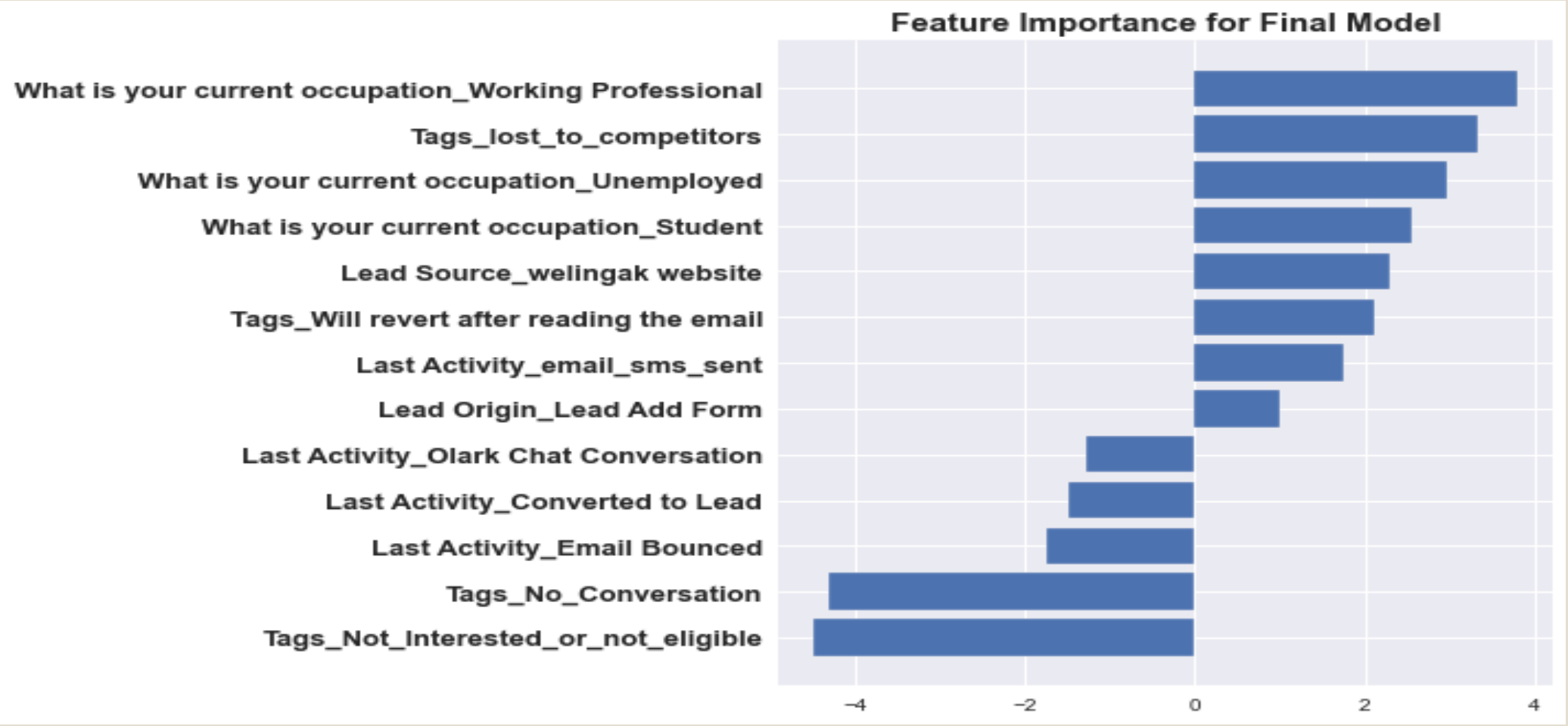


Final Model : Train ROC 0.96



Final Model: Test ROC 0.96

Final Features for the Model



Lead Score: 0-100

- **According to the probability, we find the probability score of each lead Number.**

	Lead Number	Scores
3556	626413	92
8054	588106	99
1847	641584	10
1767	642202	100
3019	630841	100
6355	601139	2
505	655122	98
6887	596751	39
4995	611994	1
6936	596447	1

Finds of Final Model:

Positive

- **Working Professional are the most prominent factor for predicting the hot leads.**
- **Tags which suggests that clients are lost to competitors should rather be focused more than ignoring them.**
- **Clients which are student or unemployed could be a potential hot lead.**
- **Leads which are generated from welingak website should be given importance and call should be made to them.**
- **Clients which are reverting to the mails are also the prominent customers.**
- **Clients which are closed by Horizons or lost to competitors have the highest conversion rate, but these clients are very limited.**

Negative

- **Customers whose last Activities were on Olark Chat should not be given much of importance.**
- **Same as for the clients which are already Converted to Lead and whose emails are bounced cause they might not be interested.**
- **Customers whose numbers are unreachable, switched off, wrong number provided or they are busy should be ignored.**
- **Customers which are interested in other courses should also be ignored.**
- **Customers who disagrees with the university should not be called.**

Other Findings:

- Google is the prominent source of getting leads but the conversion rate of it is only 40%
- House wives and working professional are more robust to conversion
- Students are converting in leads but their conversion rate is not prominent yet they could be potential lead.
- Clients which are in Service excellence has the least interest towards the given course as they are less in numbers as well as their response rate is poor than the others.
- Clients who are in management are most favorable people to opt for this course. They are good in leads as well as response rate.
- Clients who are from Banking, Investment and Insurance Domain are the most prominent leads
- Clients which are closed by Horizons or lost to competitors should not be ignored.

Recommendations

- The organization must do way with giving a free copy of "Mastering The Interview" to save cost as it has minimal effect on lead conversion.
- Leads spending more time on the website are more likely to be converted, hence the organization must work on building a more engaging website.
- It is better to text or mail the client first rather than ringing directly.
- We have poor response from African countries except south Africa.
- Western Asia and Eastern Asia countries have good response rate but we have approached less in these countries. Organization can expand its business there.

Thank You