LEAD SCORING CASE STUDY

GROUP STUDY ASSIGNMENT SUBMITED BY:-

- Pranavu G. :- drpranav12322@gmail.com
- Divya P. Banalar :- divyabanakar98@gmail.com
- Shaikh M. Hussain :- digitalstudy007@gmail.com

upGrad CS55 STUDENT GROUP

-: PROBLEM STATEMENT :-

- An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. People browse the courses or fill up a form for the course or watch some videos. They fill up a form providing their email address or phone number, also gets leads through past referrals.
- Based On Gathered data company start making calls, writing emails, etc. The typical lead conversion rate at X education is around 30%, Which is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. A typical lead conversion process can be represented using the following funnel:

- 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, you need to nurture the potential leads well in order to get a higher lead conversion.

-: WE HAVE DATA :-

- 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.
- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not

-: GOALS OF THE CASE STUDY :-

1) **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.**

2) **There are some** more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These **problems** are **provided** in a separate doc file.

Lead Conversion Process Demonstrated as a funnel

✓ **wherein 1 means it was converted and 0 means it wasn't converted.**

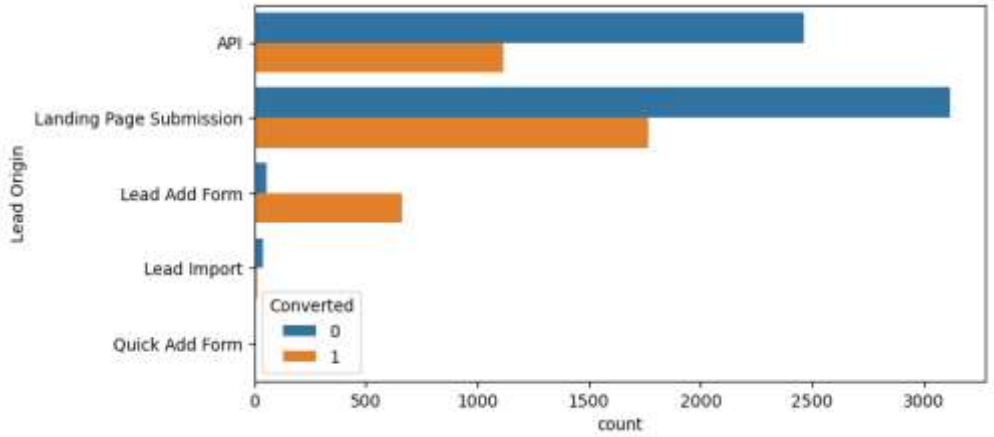
✓ **Based on the LR model we'll make our recommendations.**

-: BASED ON LEADS DATA SET :-

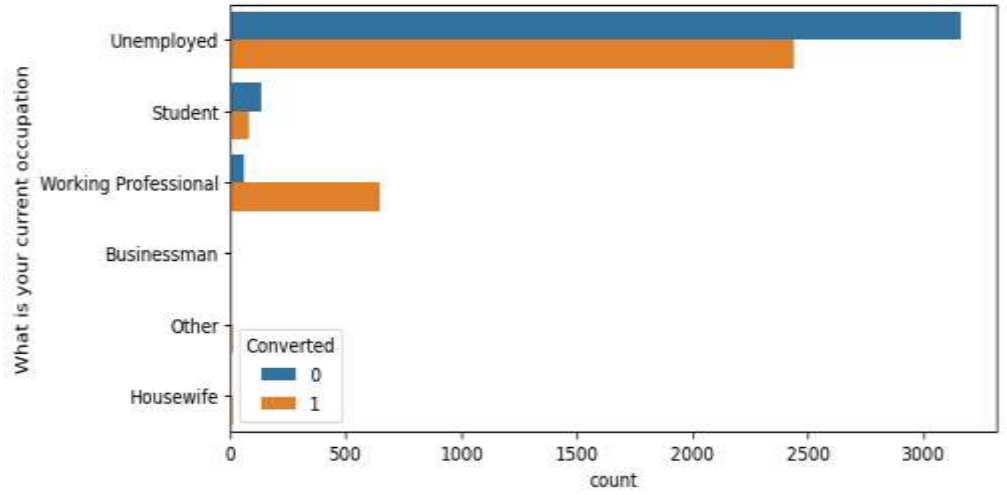
- We Found Total **9240 rows and 37 columns**, Among Them **7** columns are **numerical** variables & **30** columns are **categorical** variables. Also There are **no duplicate** values in data, But There are **some outliers present in a few columns.**

-: UNDERSTANDING THE DATA & VISUALIZATION :-

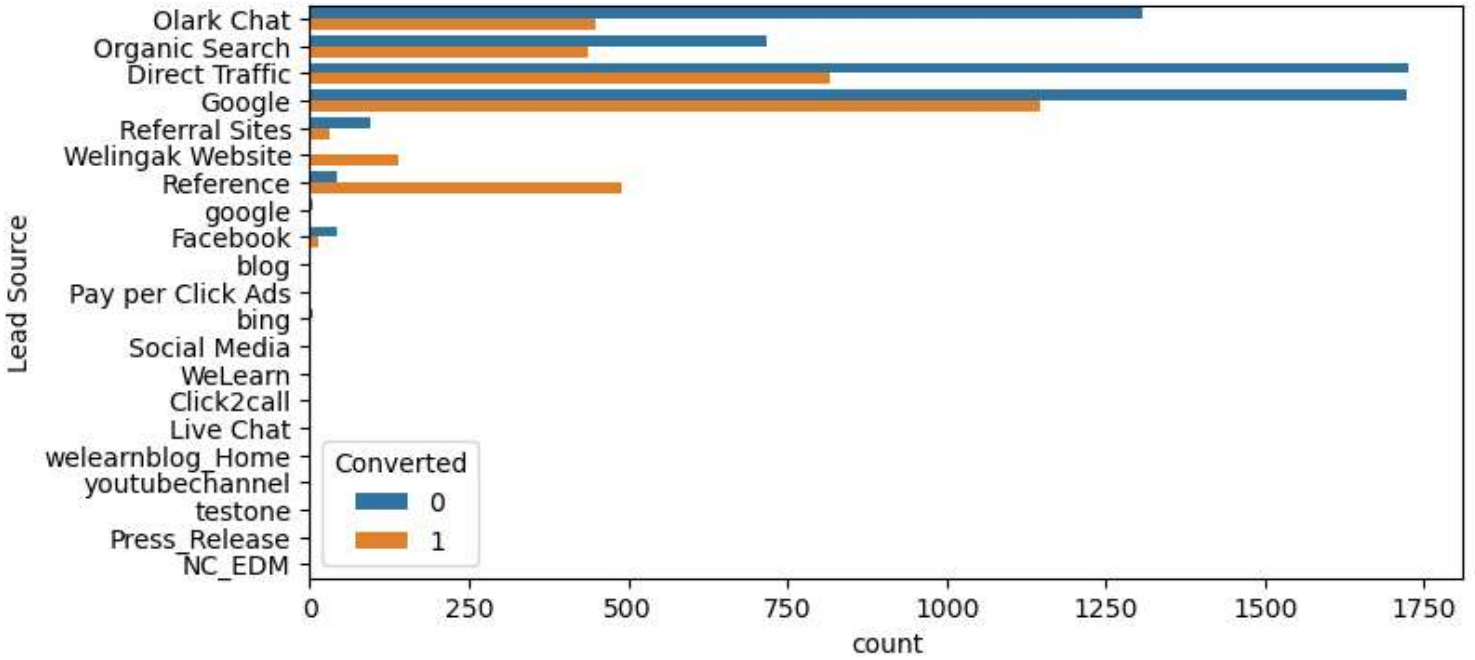
Understanding Lead Conversion and Lead Origin



Understanding Lead Conversion and Current Occupation



Understanding Lead Conversion And Lead Source



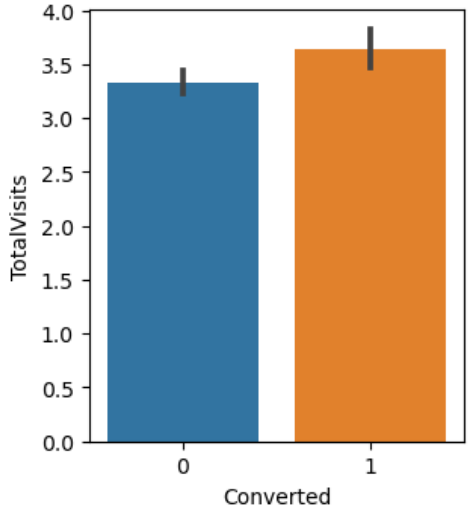
❑ **From Above DATA Analysis We Can Say That:**

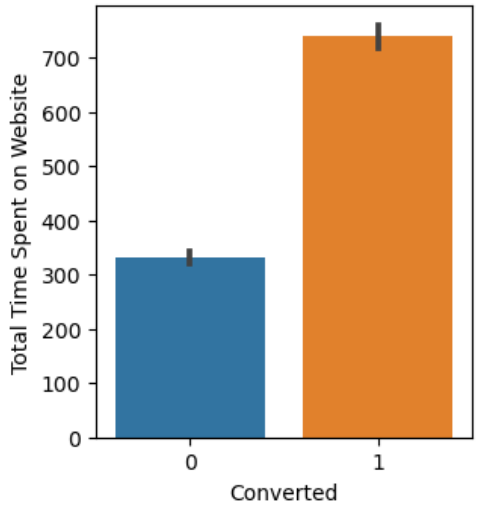
✓ lead conversion is higher from 'Landing Page Submission'

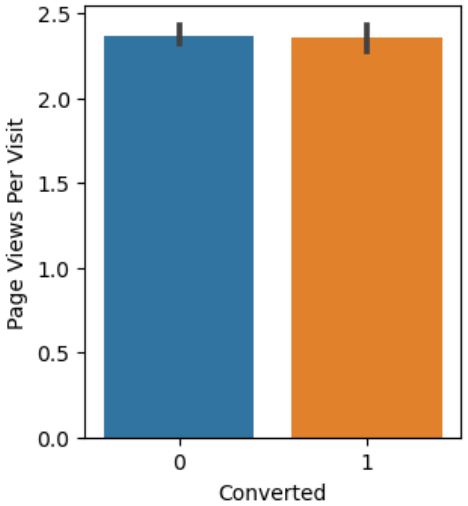
✓ Major lead conversion is from the 'Unemployed Group' followed by 'working professional'

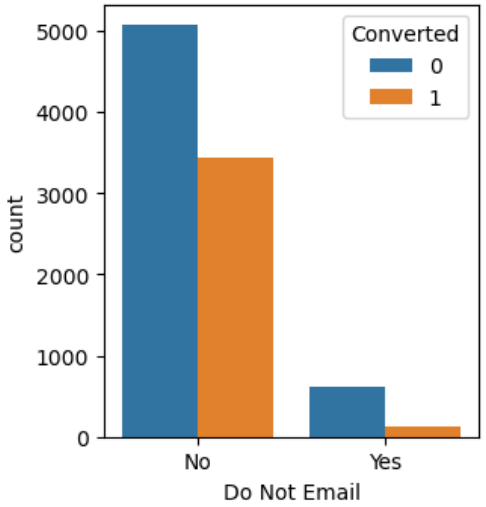
✓ Important lead conversion source is 'Google'

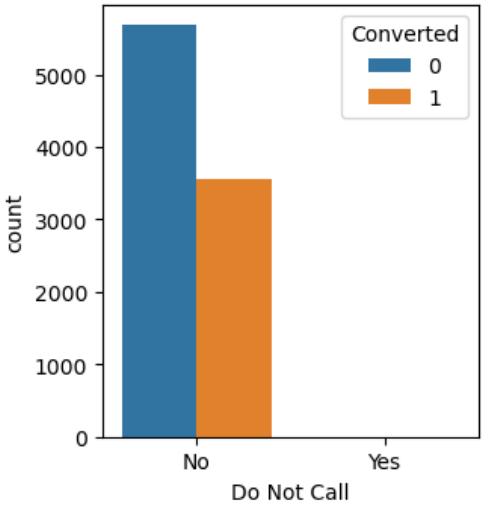
Understanding the Lead Conversion on TotalVisits, Total Time Spent on Website, Page Views Per Visit, do not email, do not call.









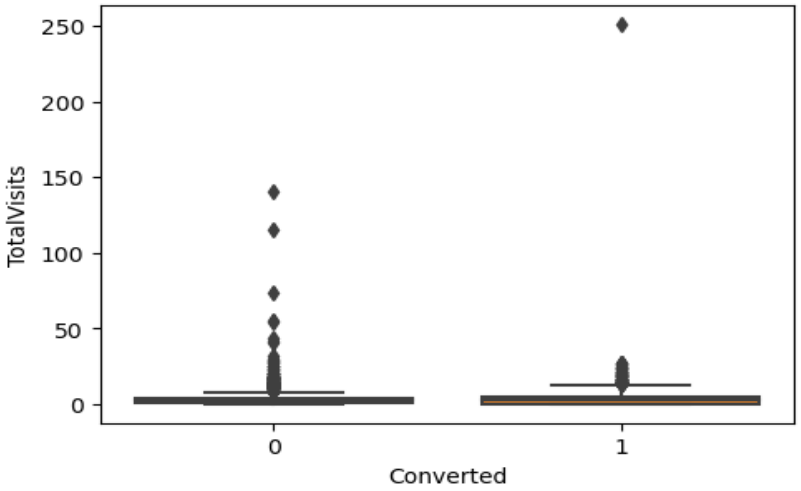


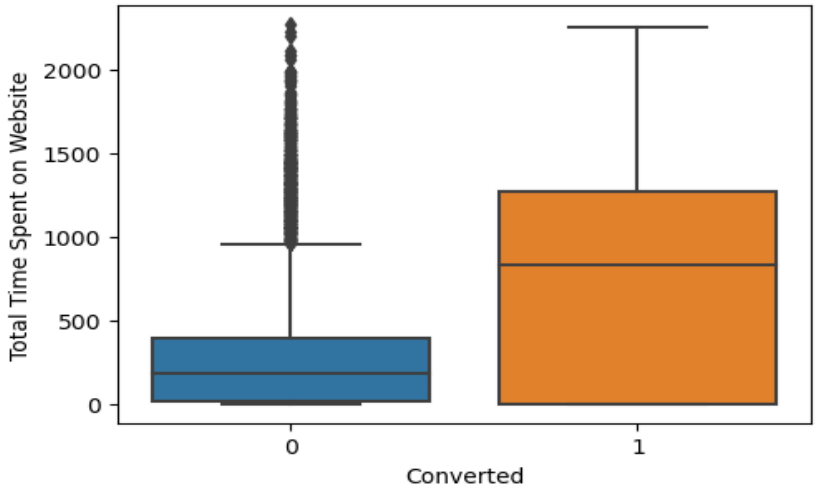
❑ From Above DATA Analysis We Can Say That:

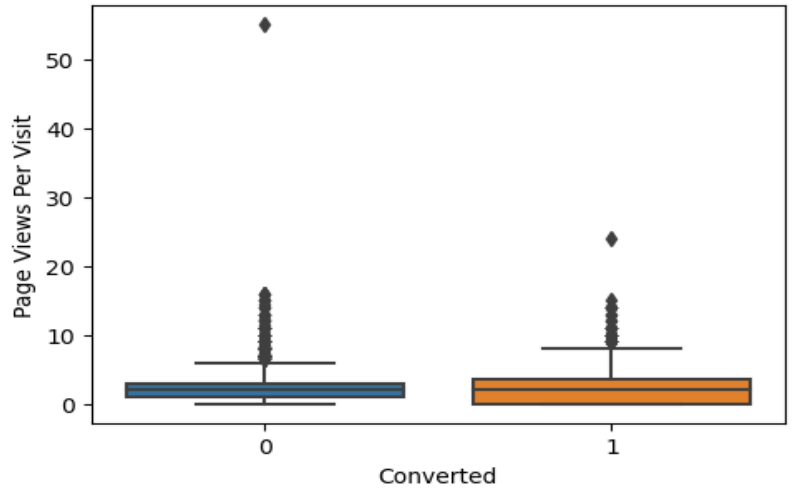
✓ Major lead conversion happened from 'TotalVisits, Total Time Spent on Website, Page Views Per Visit'

✓ Lead conversion has happened through 'Do Not Email' compared to Do Not Call'

 # checking for outliers in Numerical variables:





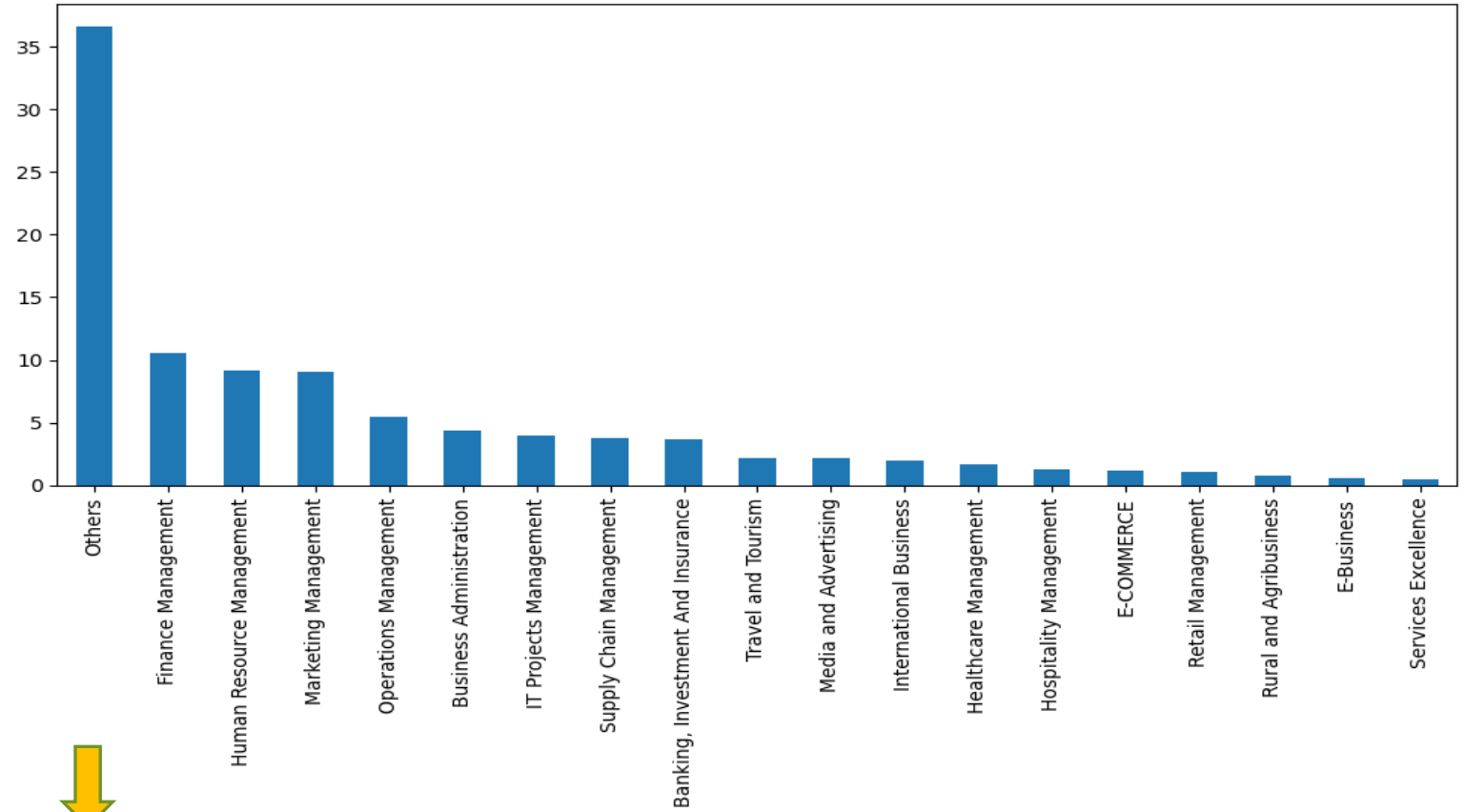


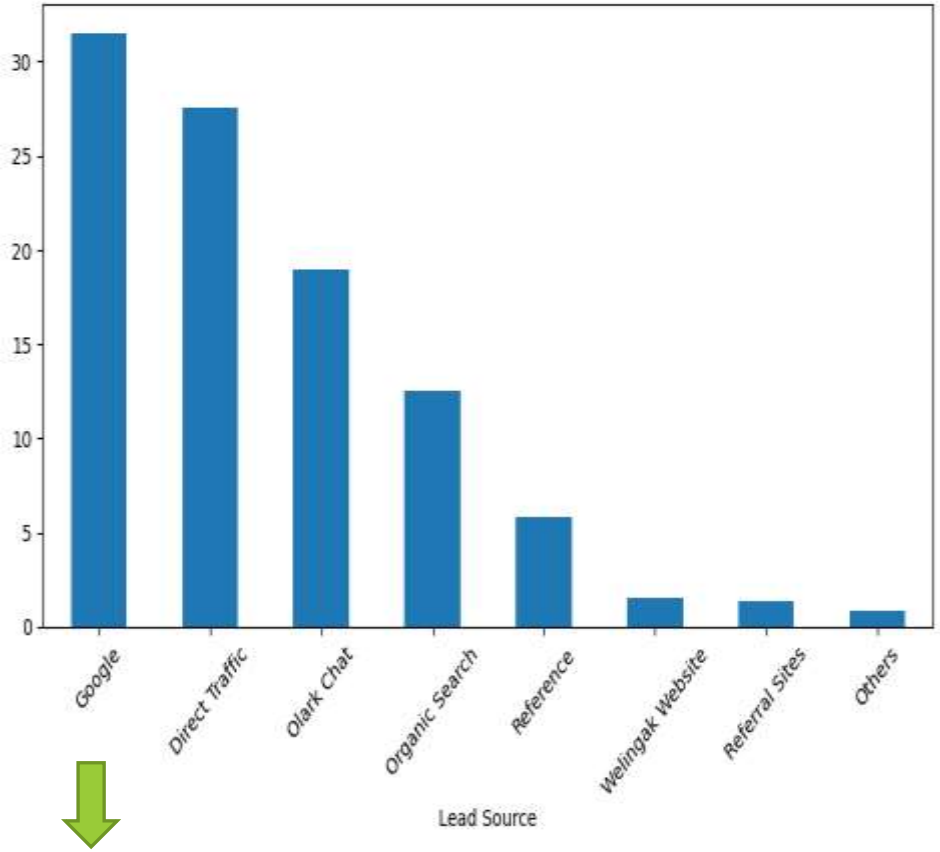
✓ **There are potential outliers present in 'total visits' and 'pages views per visit'**

-: DATA CLEANING :-

- Cleaning the dataset by dropping the redundant variables/features Also Replacing 'Select' Value With NaN Value
- After dropping null value Found Total No. of columns > 40% are 21.
- We confirmed that the null value columns > 40% are dropped and Based on that Missing Values Imputation is for the columns: 'Last Activity', 'Tags' are provided to sales team. **NOTE:-** We will remove them before model building as the we don't a model having these features

-: MISSING VALUES IMPUTATION :-

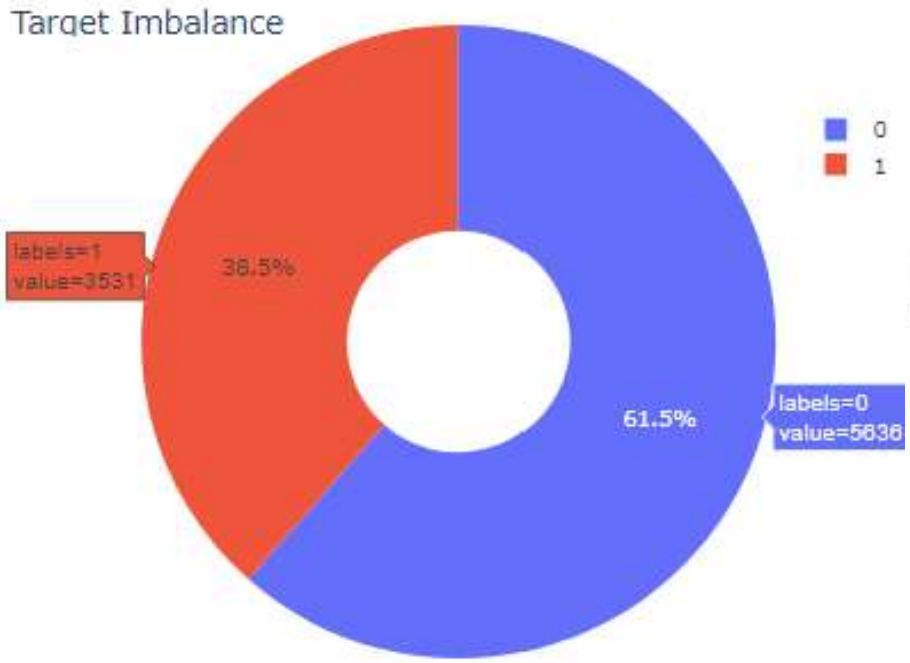


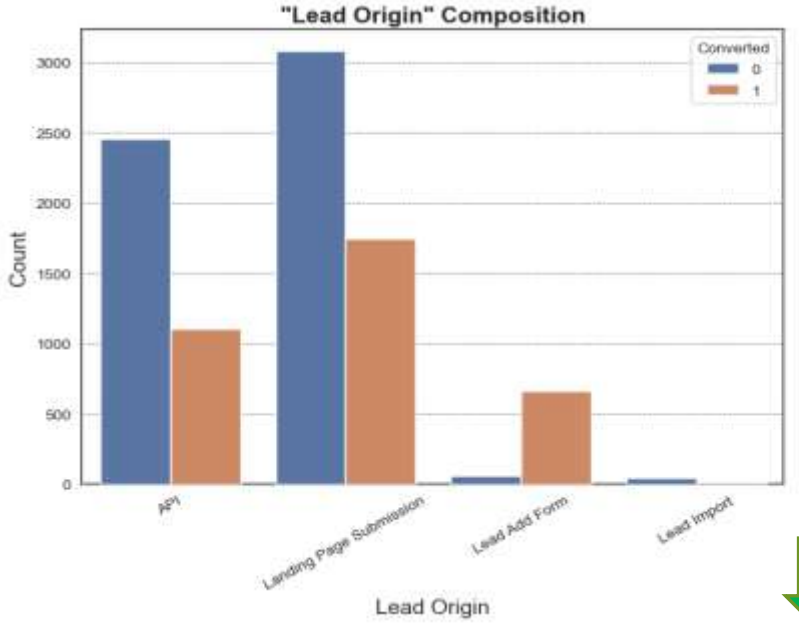


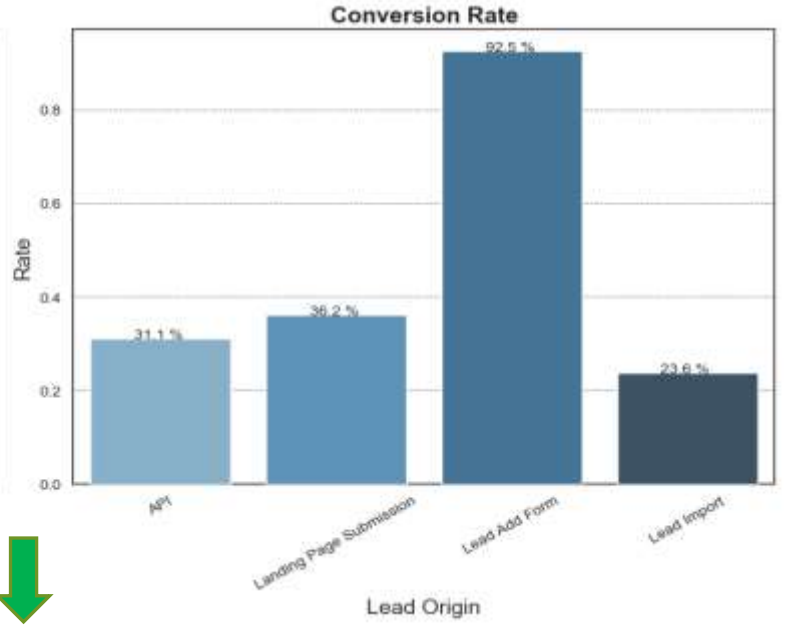
 # Replacing NaN's with 'Others' as these values may correspond to students/freshers

 # Handling 'Lead Source' By Imputation of most frequent level of 'Google' (mode - imputation) with replacing google By Google & combining less frequent levels into one, 'Others'

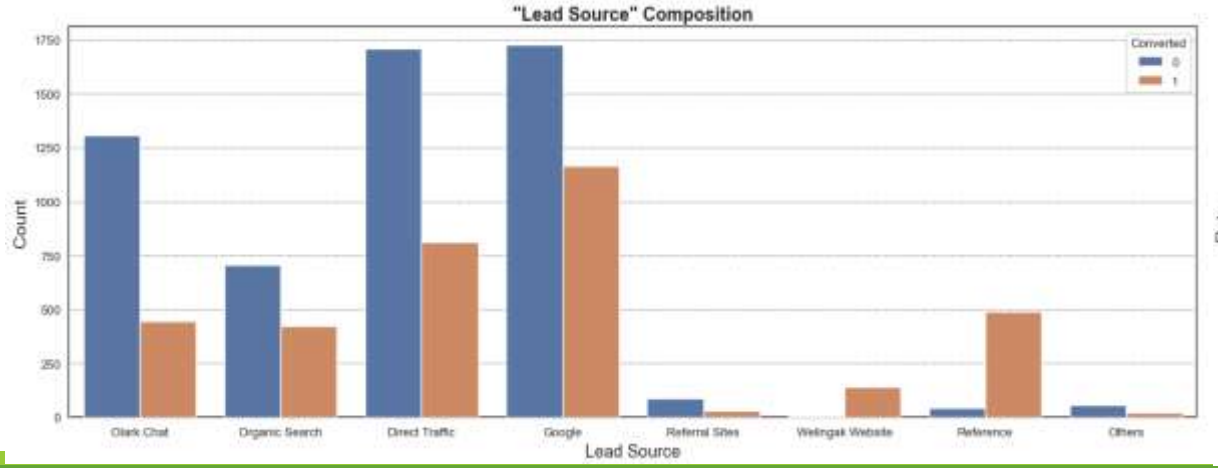
-: EXPLORATORY DATA ANALYSIS :-

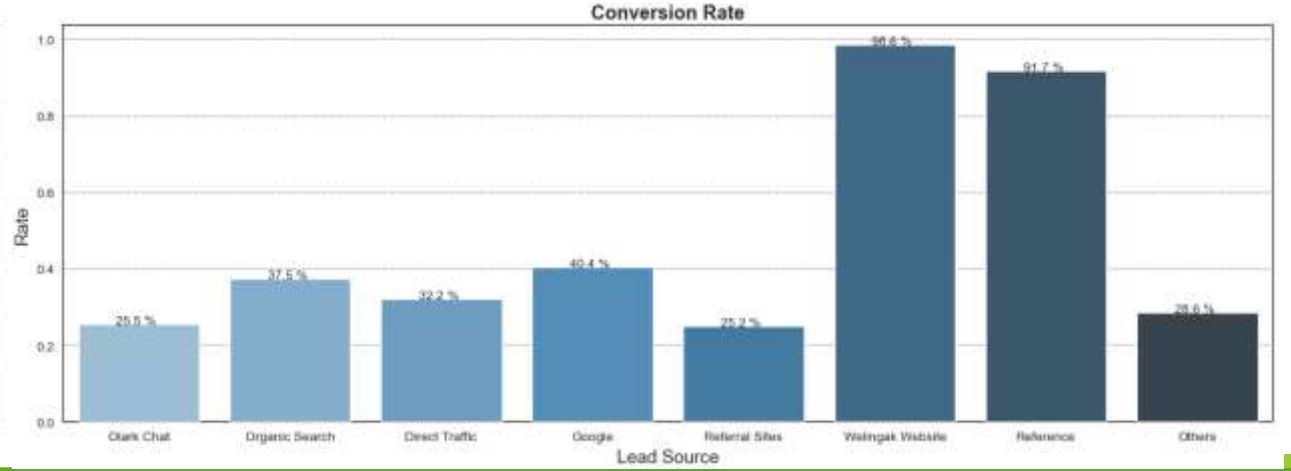




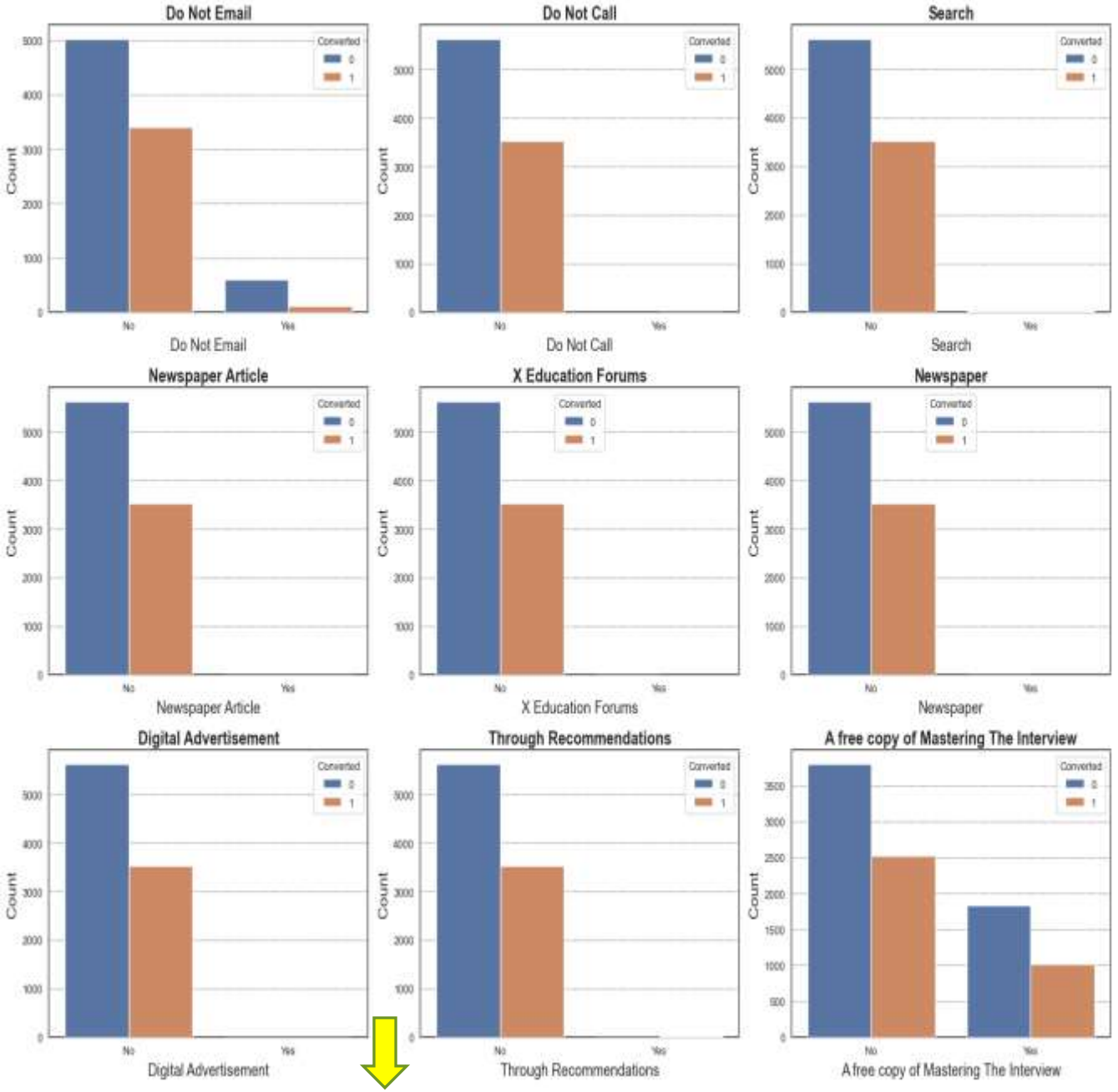


- ✓ No. of customers - identified as leads by API and Landing Page Submission are highest, but their conversion rate is less than the average **overall conversion rate ~ 38.5 %.** Even though, Lead Add Form identifies brings in less leads but the **conversion rate** of the leads identified by the it **is very high**. **Company should try to bring in more leads by 'Lead Add Form'**



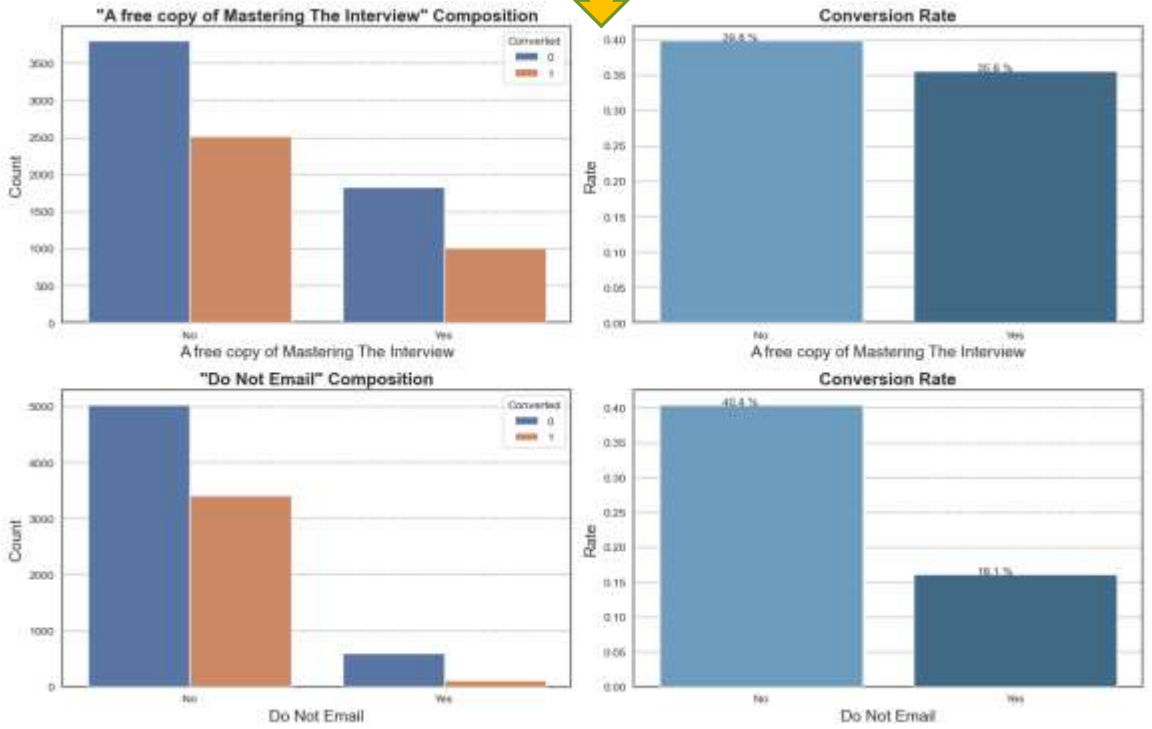


- ✓ **Highest Number of leads come from Google and Direct Traffic.**
- ✓ **Conversion rate of leads from direct traffic is less than overall**
- ✓ **High % of leads from welingak website & References converted.**
- ✓ **Invest more resources into acquiring leads from these sources.**



visualizing binary categorical variables

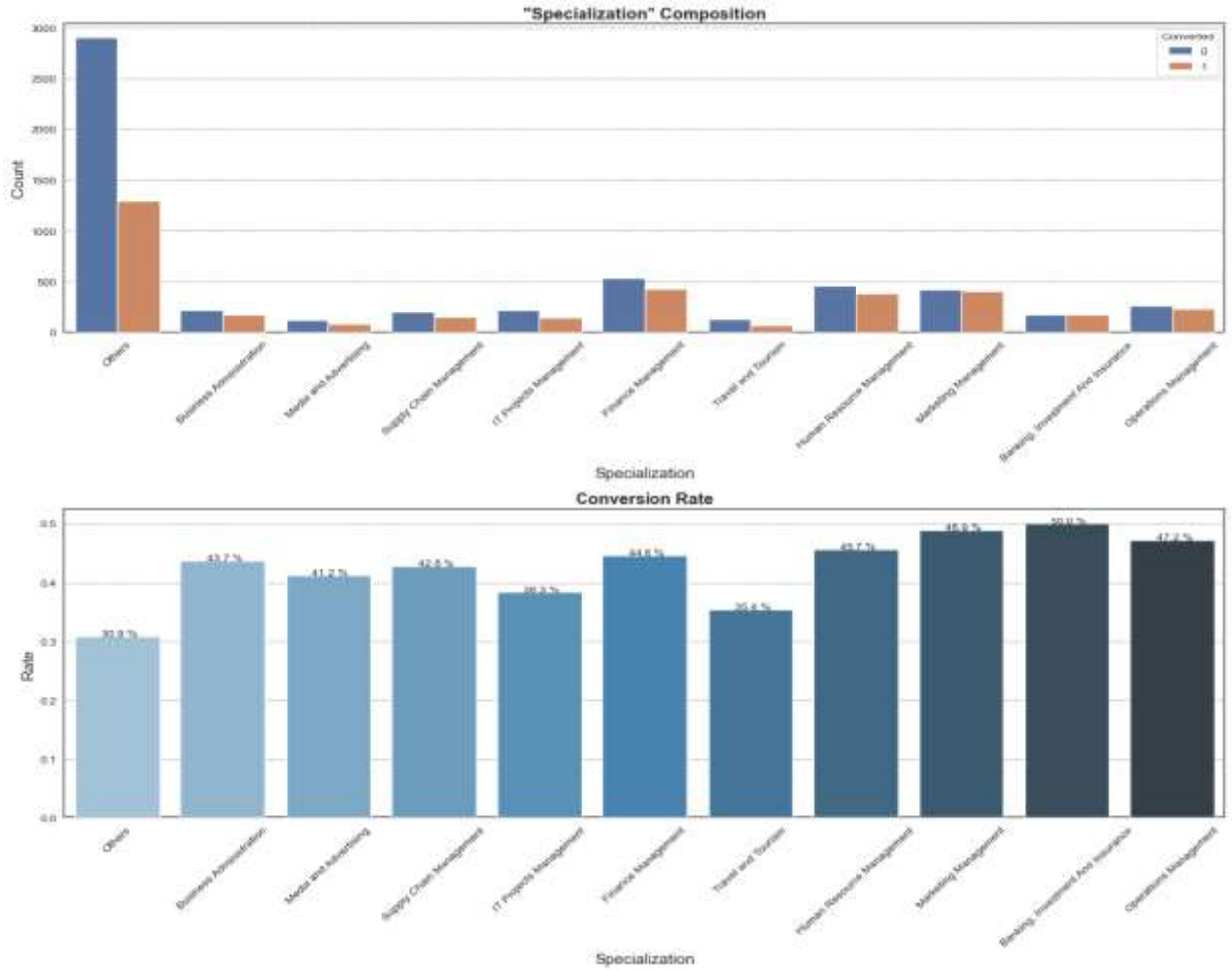
For better clarity after dropping: Again visualizing binary categorical variables



- ✓ Customers who selected to 'Do not get email' about course converted significantly more than those who selected to get emailed about course. This group can be targeted more.
- ✓ Customers who did not want a copy of 'Mastering The Interview' had slightly more conversion rate

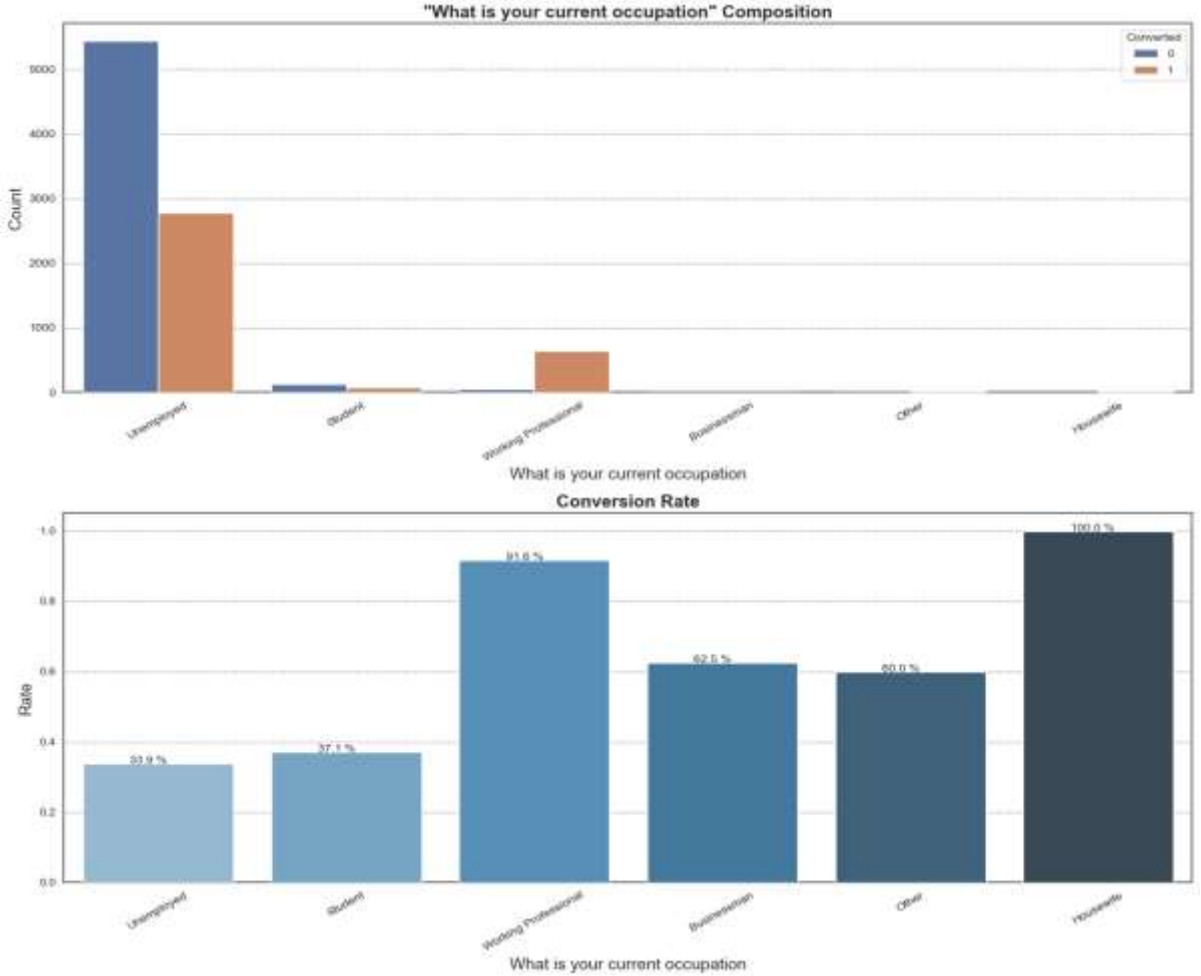
- ✓ It looks like for many of the variable here, one of the level is highly dominant and the other has almost no contribution.
- ✓ These variable are not useful for the model, so we can drop them.

Rate of conversion On Specialization



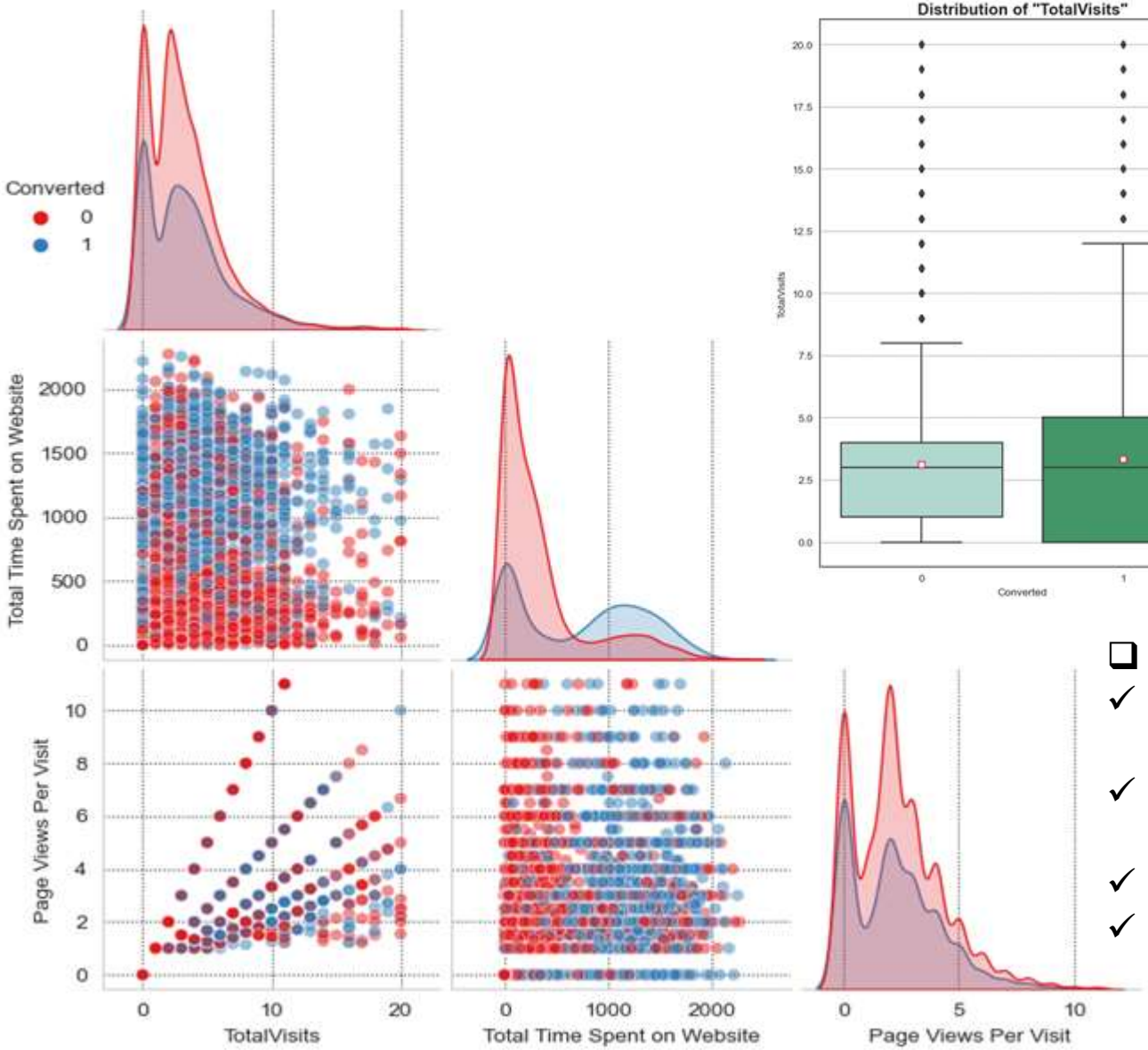
- ✓ Leads from Management sector, like HR and Marketing Management, and Banking, Investment and Insurance specialization are relatively more likely to convert.
- ✓ Their average conversion rate is higher than the overall average.
- ✓ These groups can be targeted more

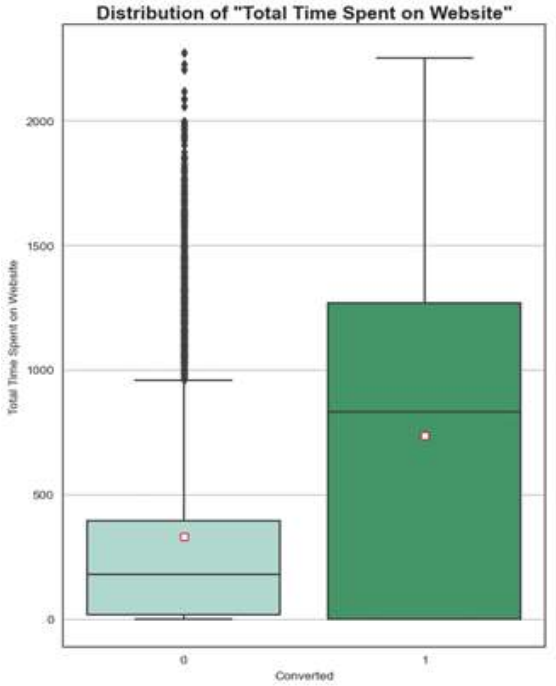
Rate of conversion On Current occupation

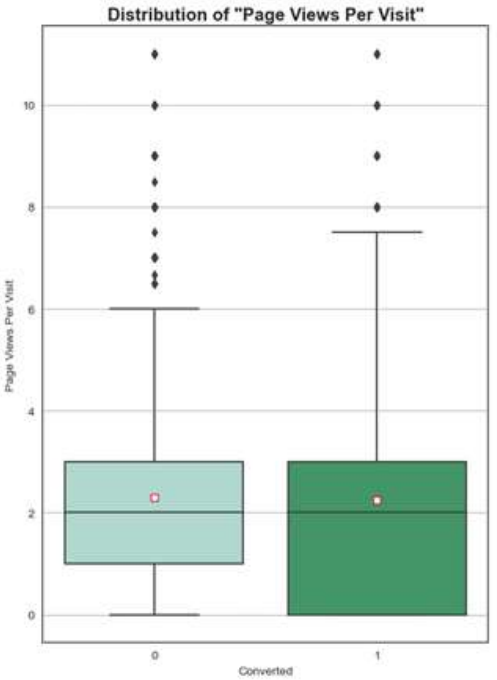


- ✓ Unemployed people are least likely to convert. Working professionals have a very high conversion rate. They should be targeted more.
- ✓ Housewives have a 100% conversion rate but the data for housewives is too small to make conclusive decision

-: ANALYSING NUMERICAL VARIABLES :-







❑ **Outcomes from Numerical analysis:**

- ✓ **Average and median total number of visits made by the leads are almost the same.**
- ✓ **The average and median total number of page views per visit are also almost the same.**
- ✓ **The values are more spread out for the leads that converted.**
- ✓ **Leads that spend more time on the website are more likely to convert. These people show be approached more. Also, websites can be made more engaging and user-friendly to improve the numbers.**

-: BINNING :-

- After creating bins we removed the outliers and are now good to go. Before creating the dummy variables let's remove redundant columns/variables.

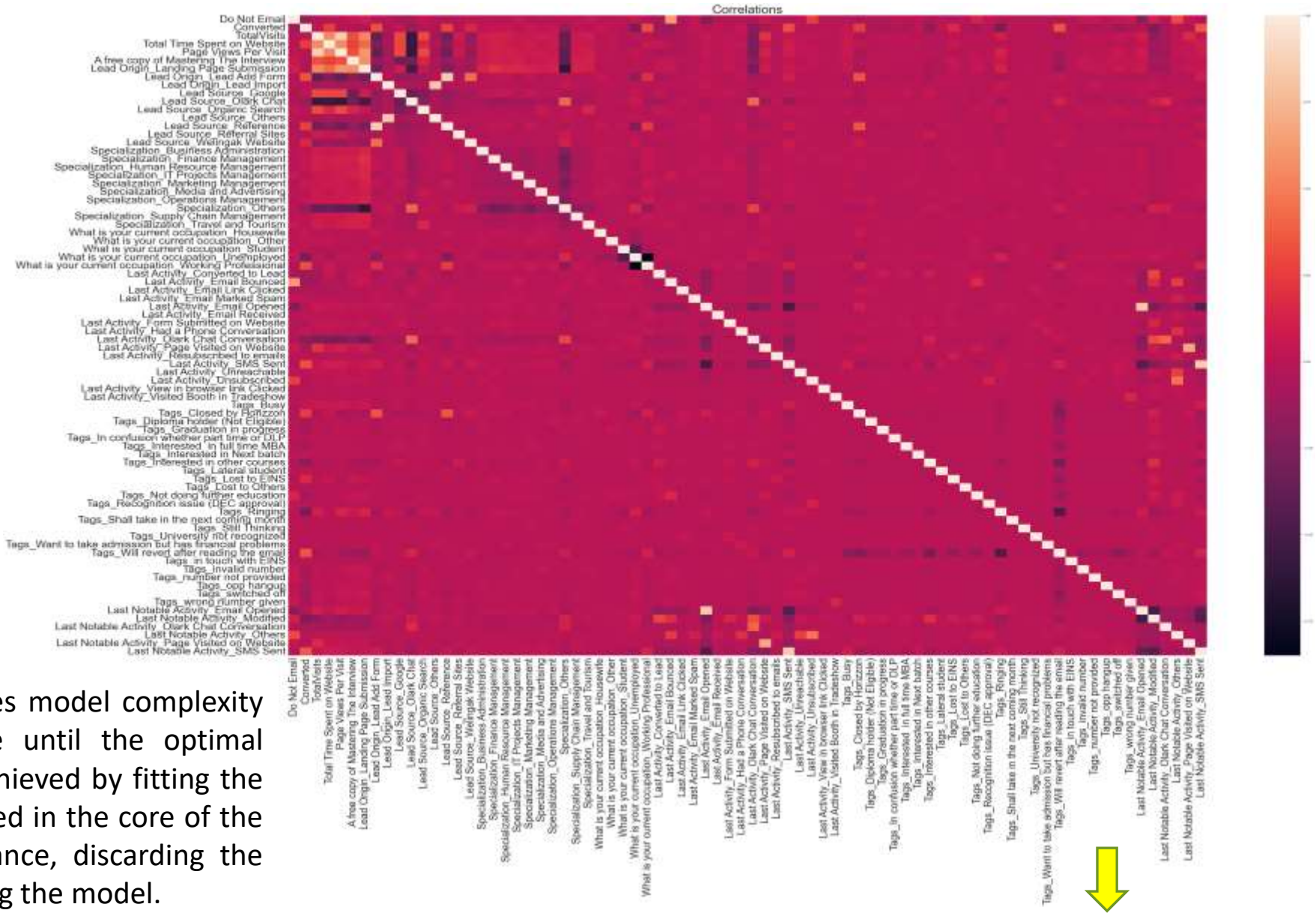
- Also from above we know columns : 'Last Activity', 'Tags', 'Last Notable Activity' activity columns came from sales team, thus we will drop these redundant columns.

-: MODEL-BUILDING :-

❑ **RFE (coarse tuning)**

✓ Recursive feature elimination reduces model complexity by removing features one by one until the optimal number of features is left. This is achieved by fitting the given machine learning algorithm used in the core of the model, ranking features by importance, discarding the least important features, and re-fitting the model.

-: DATA PREPARATION - CORRELATION IN THE DATASET :-



✓ **As observed, there are high correlations between some of the variables and therefore, data has multicollinearity.**

-: IN TRAIN TEST DATA 4TH FINAL MODEL:-

Generalized Linear Model Regression Results

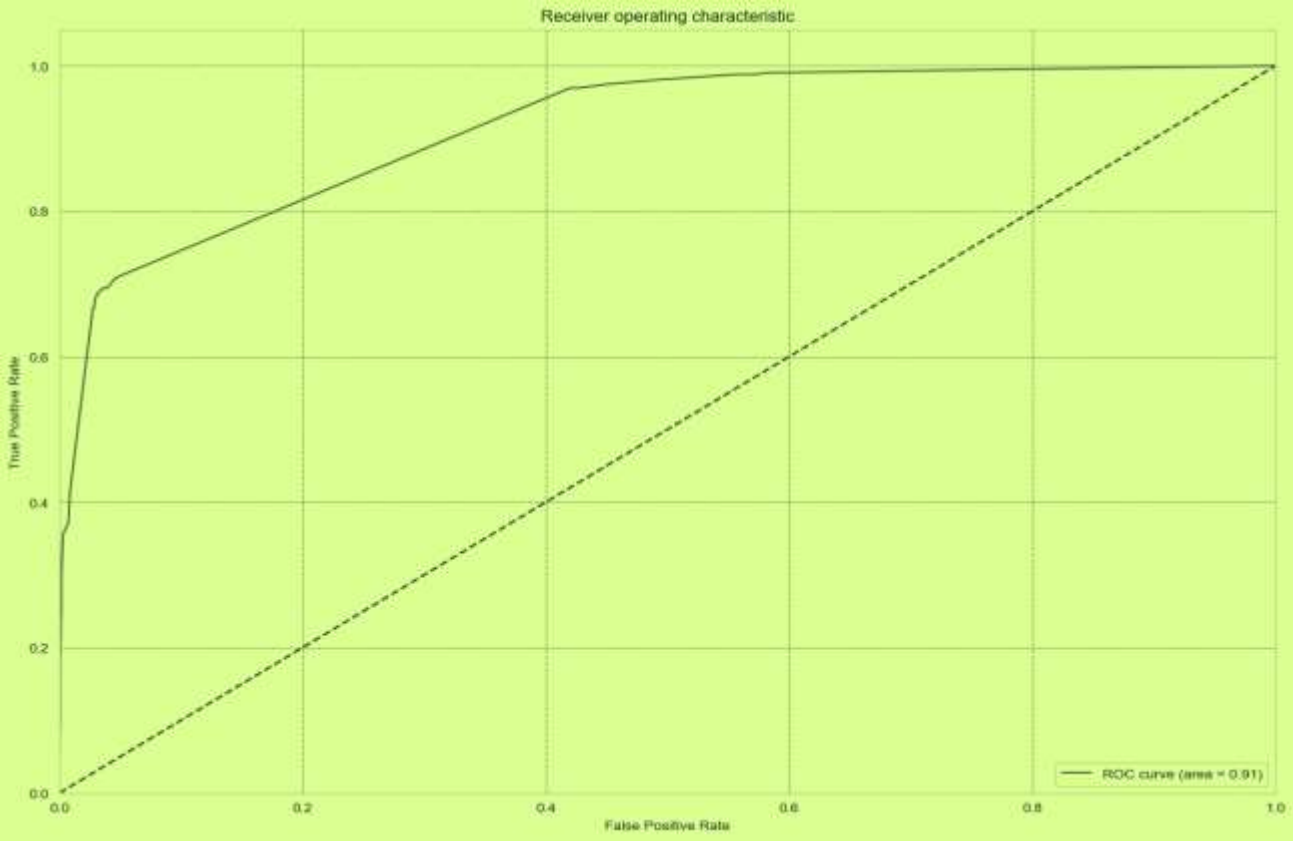
Dep. Variable:	Converted	**No. Observations:**	6416
Model:	GLM	**Df Residuals:**	6402
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2103.6
Date:	Sat, 14 Oct 2023	**Deviance:**	4207.2
Time:	14:22:57	**Pearson chi2:**	1.03e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.4891
Covariance Type:	nonrobust		

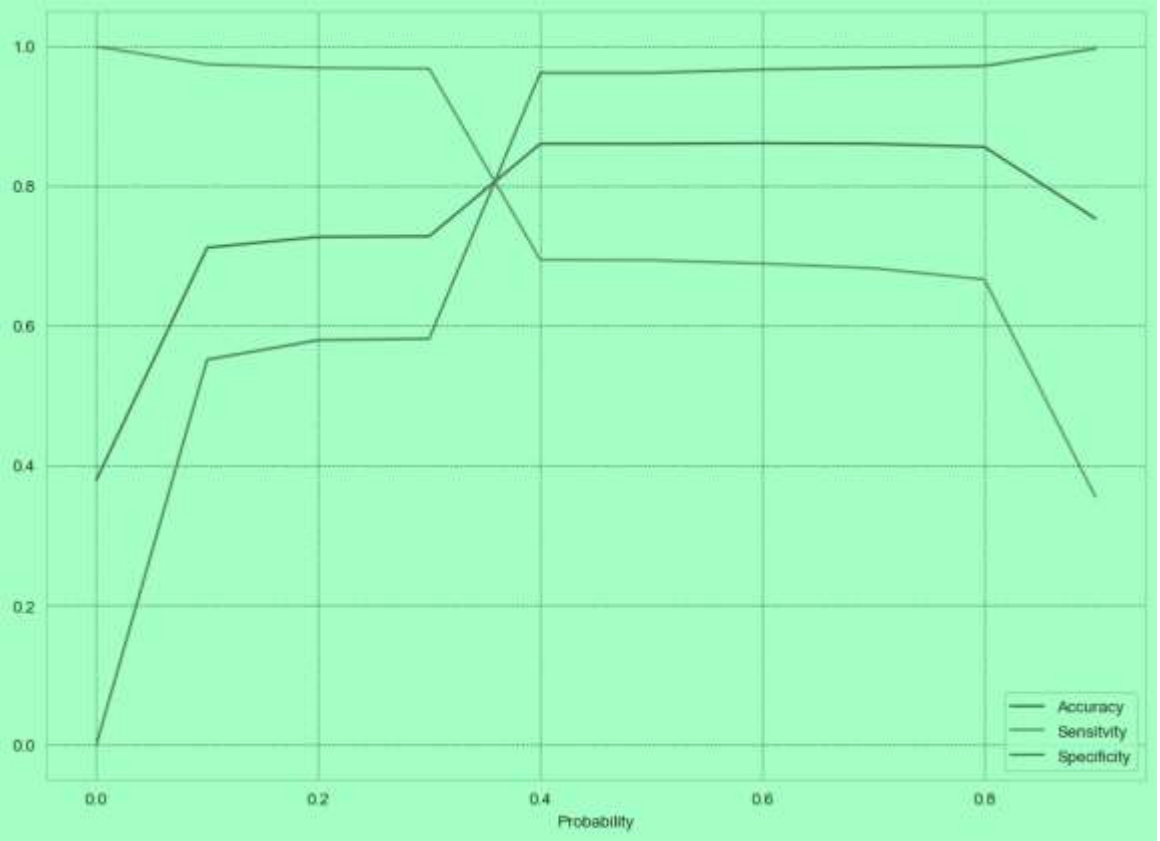
	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8324	0.299	-9.482	0.000	-3.418	-2.247
Do Not Email	-1.5141	0.179	-8.474	0.000	-1.864	-1.164
Lead Origin_Lead Add Form	2.7735	0.225	12.346	0.000	2.333	3.214
What is your current occupation_Unemployed	-1.8150	0.276	-6.585	0.000	-2.355	-1.275
What is your current occupation_Working Professional	1.3588	0.355	3.825	0.000	0.663	2.055
Last Activity_Had a Phone Conversation	2.3144	0.775	2.987	0.003	0.796	3.833
Tags_Busy	3.9907	0.289	13.790	0.000	3.423	4.558
Tags_Closed by Horizzon	8.4588	0.741	11.416	0.000	7.007	9.911
Tags_Lost to EINS	8.5367	0.743	11.486	0.000	7.080	9.993
Tags_Ringing	-0.7481	0.297	-2.516	0.012	-1.331	-0.165
Tags_Want to take admission but has financial problems	3.3828	1.245	2.717	0.007	0.943	5.823
Tags_Will revert after reading the email	3.8682	0.197	19.596	0.000	3.481	4.255
Tags_in touch with EINS	3.3947	0.826	4.110	0.000	1.776	5.013
Last Notable Activity_SMS Sent	2.6687	0.107	24.845	0.000	2.458	2.879

Features	VIF
What is your current occupation_Unemployed	4.54
Tags_Will revert after reading the email	3.89
Tags_Ringing	1.69
What is your current occupation_Working Profes...	1.50
Last Notable Activity_SMS Sent	1.44
Tags_Closed by Horizzon	1.37
Lead Origin_Lead Add Form	1.30
Tags_Busy	1.12
Do Not Email	1.10
Tags_Lost to EINS	1.10
Last Activity_Had a Phone Conversation	1.02
Tags_Want to take admission but has financial ...	1.02
Tags_in touch with EINS	1.01

- ✓ P-value of all the predictor variables are less than < 0.05 ; indicates - statistically significant
- ✓ VIF values less than 5, it suggests that there is no strong multicollinearity among the independent variables in regression model. This is a good sign because high multicollinearity can make it challenging to interpret the impact of individual predictors.
- ✓ Hence, we can consider this as our final trained model

-: PREDICTION ON TRAIN DATA SET :-

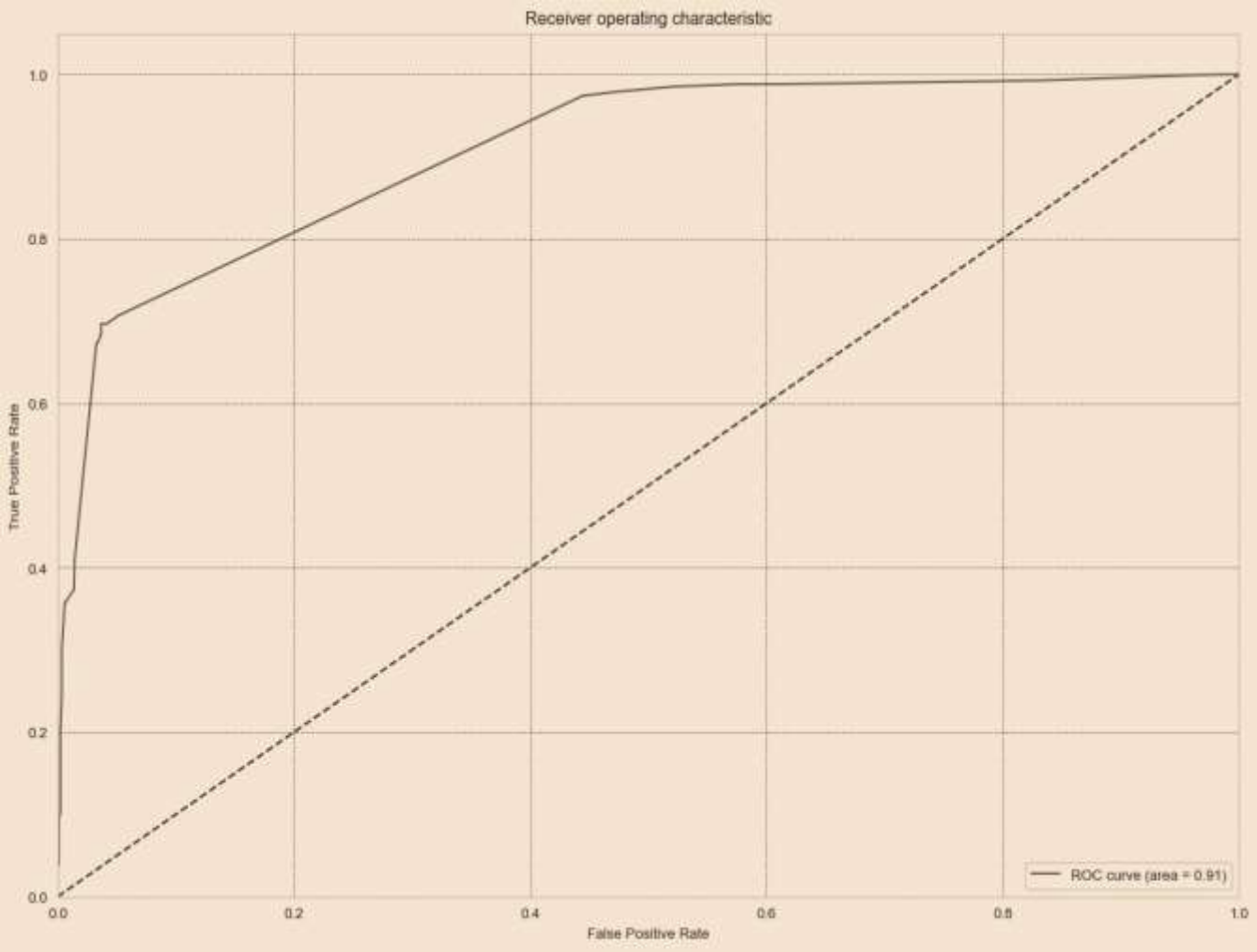




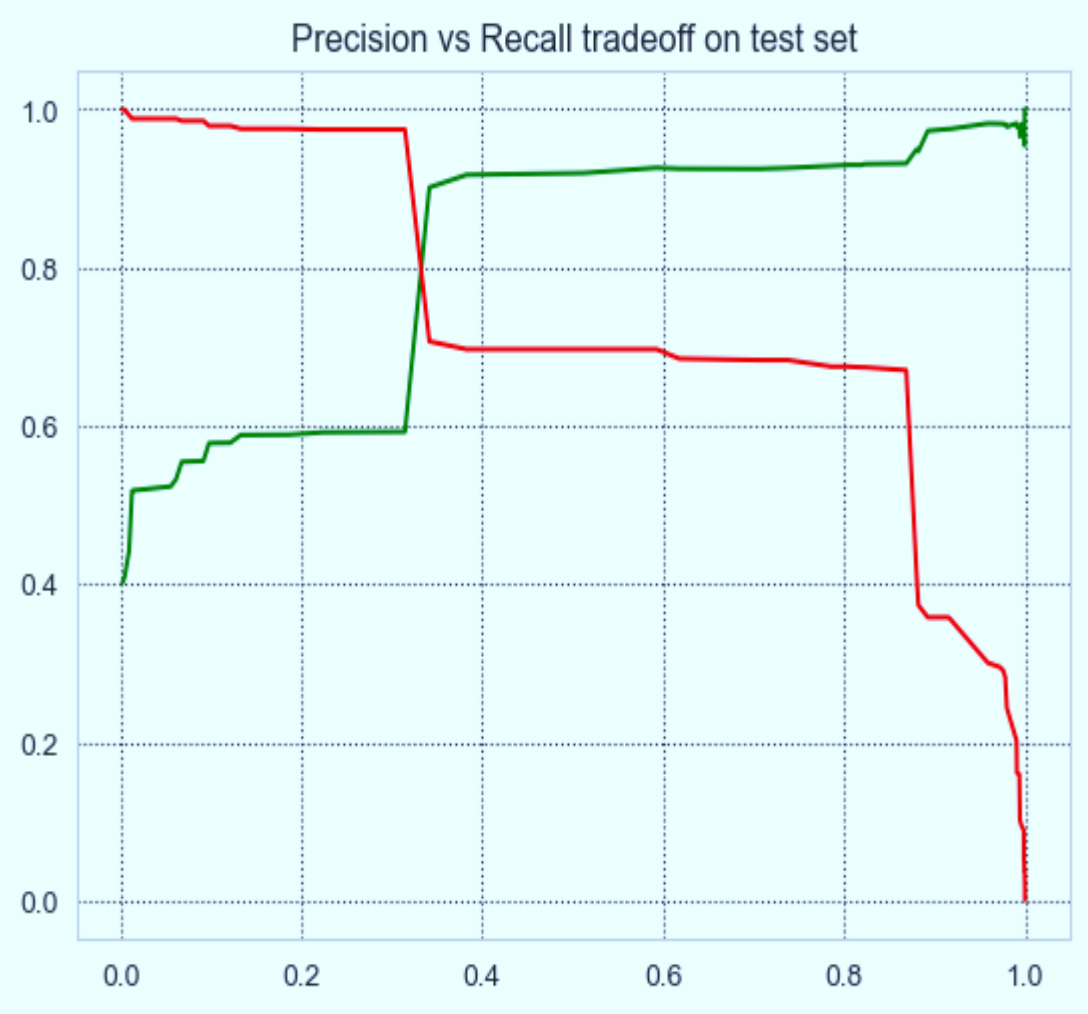
❑ ROC Curve Plotting:

- ✓ ROC curve shows the trade off between True positive rate and False positive rate - means if sensitivity increases specificity will decrease
- ✓ The curve closer to the left side border then right side of the border is more accurate
- ✓ The curve closer to the 45-degree diagonal of the ROC space is less accurate
- ✓ We would like both sensitivity and specificity to be high
- ✓ At the same time we also would not want to spend much resources on negatives(leads who will not convert)
- ✓ let's look at confusion matrix with threshold 0.4 for now.

-: PREDICTION ON TEST DATA SET :-



-: MODEL EVALUATION :-



-: VALUABLE INSIGHTS :-

```
Tags_Lost to EINS                                          8.536723
Tags_Closed by Horizzon                                    8.458760
Tags_Busy                                                  3.990670
Tags_Will revert after reading the email                   3.868249
Tags_in touch with EINS                                    3.394666
Tags_Want to take admission but has financial problems     3.382832
Lead Origin_Lead Add Form                                  2.773458
Last Notable Activity_SMS Sent                             2.668656
Last Activity_Had a Phone Conversation                     2.314417
What is your current occupation_Working Professional       1.358837
Tags_Ringing                                              -0.748086
Do Not Email                                              -1.514073
What is your current occupation_Unemployed                -1.815022
const                                                     -2.832407
dtype: float64
```

- **Most important variables to consider:**

✓ 'Tags_Lost to EINS;': If this variable is True or 1, then the log-odds increases by 8.53

✓ 'Tags_Closed by Horizzon': If this variable is True or 1, then the log-odds increases by 8.45

✓ 'Tags_Busy': If this variable is True or 1, then the log-odds increases by 3.99

✓ 'Occupation_Unemployed': If the current status 'Unemployed', then the log odds decrease by 1.81

✓ 'Do not email': If this variable is True or 1, then the log-odds decrease by 1.51

✓ 'Tags_Ringing': If the current status / tag is 'Ringing', then the log odds decrease by 0.74

-: COMPARISON BETWEEN TRAIN DATA SET AND TEST DATA SET :-

```
Train Data Set metrics:

Sensitivity: 69.46
Specificity: 96.23
Precision: 91.85
Recall: 69.46
Accuracy: 86.08
F1_Score:  79.10133395740698

Test Data Set metrics:

Sensitivity: 69.67
Specificity: 95.95
Precision: 91.95
Recall: 69.67
Accuracy: 85.46
F1_Score:  79.27461139896373
```

Evaluation metric	Train Data Set	Test Data Set
Sensitivity	69.46	69.67
Specificity	96.23	95.95
Precision	91.85	91.95
Recall	69.46	69.67
Accuracy	86.08	85.46
F1 Score	79.1013	79.2746

- **CONCLUSION:**

✓ The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost similar to train set.

✓ No concerns w.r.t Over - fitting. Model is in stable state.

✓ However, in terms of exactly predicting the lead conversion it can still do better. Since, the recall or sensitivity slightly in lower end

- **The higher the magnitude of coefficients,|Coef_|, the greater is it's control on the final decision. A high value of a variable with high positive coefficient increases the lead score while a high values of a variable with negative coefficient decreases the lead score.**

 THANK YOU