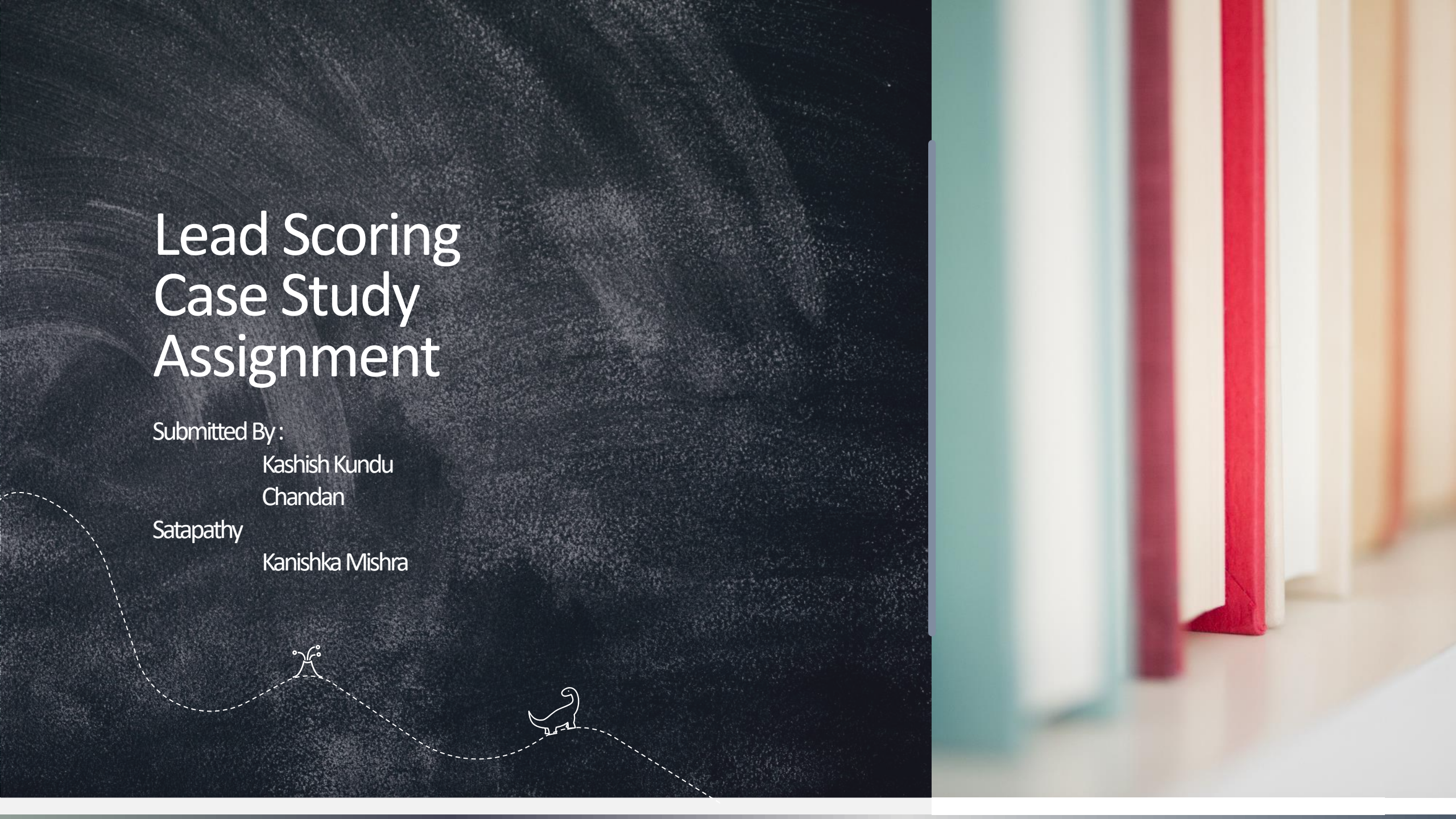
Lead Scoring Case Study Assignment

Submitted By :

Kashish Kundu

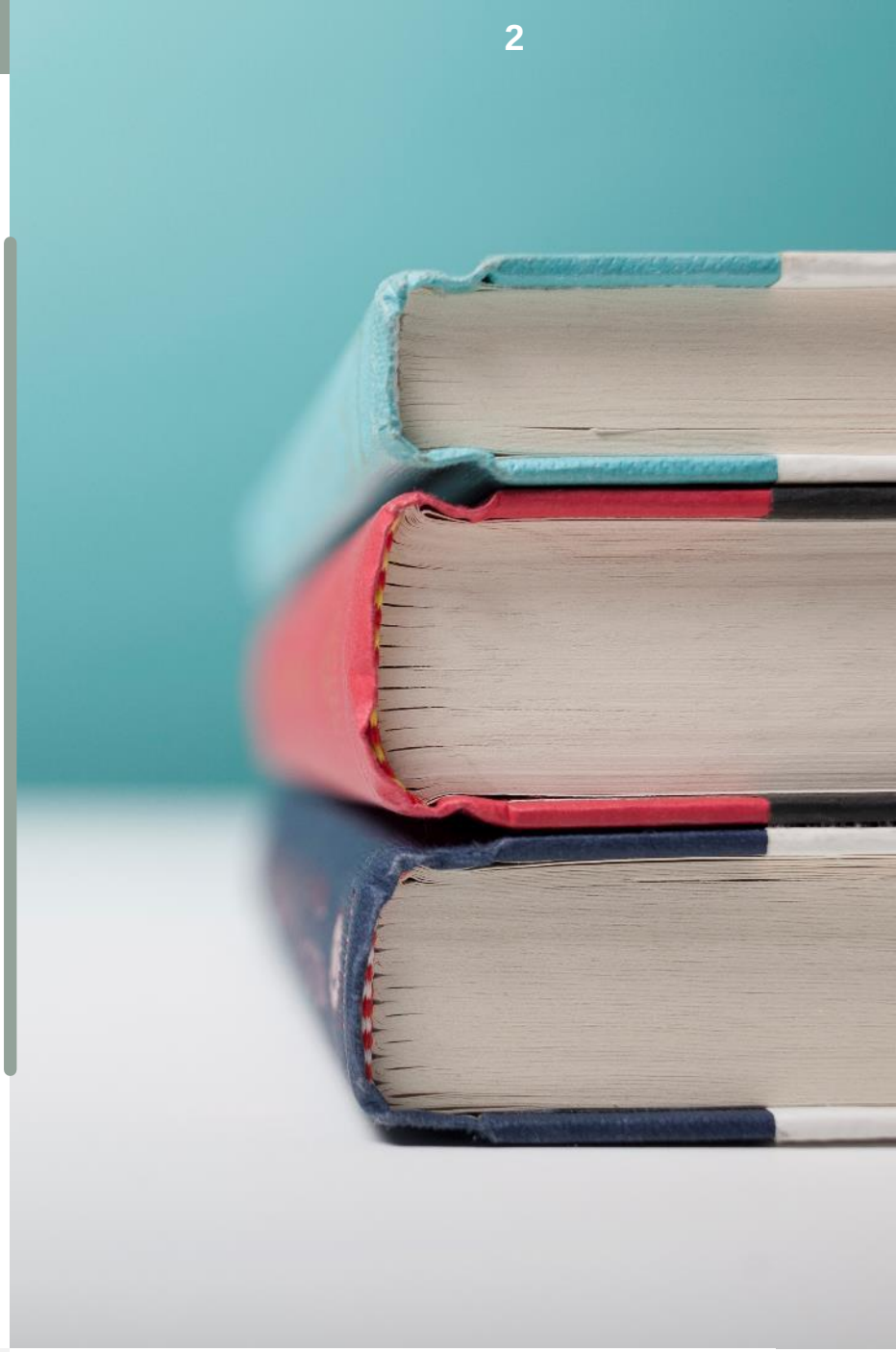
Chandan Satapathy

Kanishka Mishra



Problem statement

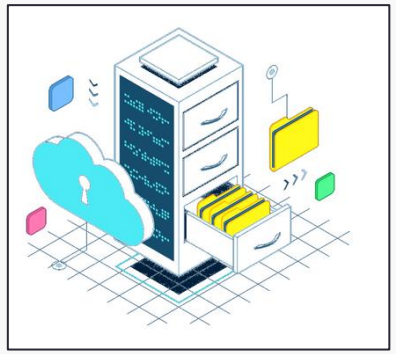
- **Challenge and Goal:** X Education faces a low 30% lead conversion rate, prompting the goal to boost it to 80% by concentrating sales efforts on high-potential leads.
- **Lead Conversion Initiative:** Employing a predictive model, X Education plans to assign lead scores, prioritizing leads for efficient resource allocation and significantly improving the conversion rate.
- **Lead Conversion Funnel:** The company visualizes the lead conversion process as a funnel, emphasizing the need to streamline and optimize the journey from lead generation to successful customer conversion.
- **Overall Objective:** X Education's overarching objective is to develop a lead scoring model that identifies and prioritizes leads with the highest conversion potential, aiming to enhance overall sales effectiveness and achieve the targeted 80% lead conversion rate.

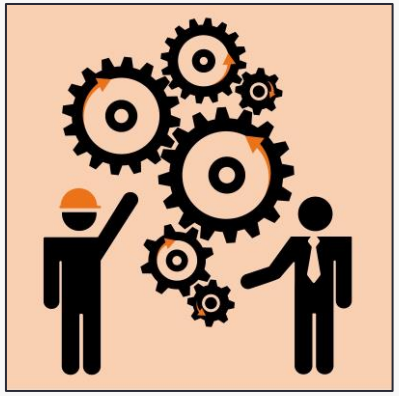


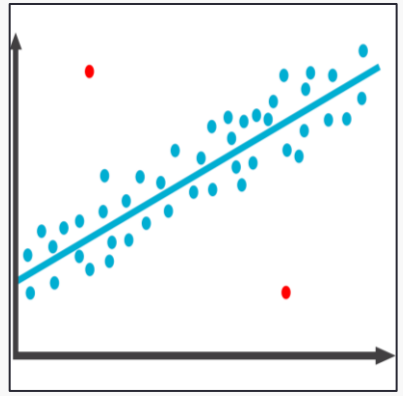
Aim

- **Logistic Regression Model for Lead Scoring:** Developed a logistic regression model to assign lead scores (0-100), aiding X Education in targeting potential leads. Higher scores indicate hot leads with a high conversion likelihood, while lower scores denote colder leads less likely to convert.

- **Model Adaptability:** Ensured the logistic regression model's adaptability to future changes by addressing additional company problems outlined in a separate document. Recommendations and model adjustments will be presented in the final PowerPoint to optimize lead conversion strategies.

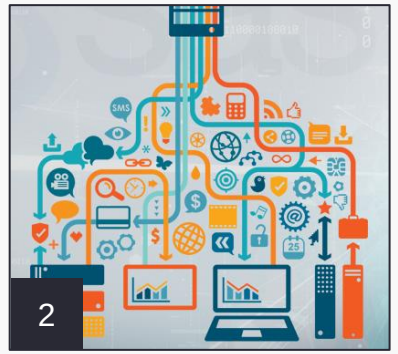


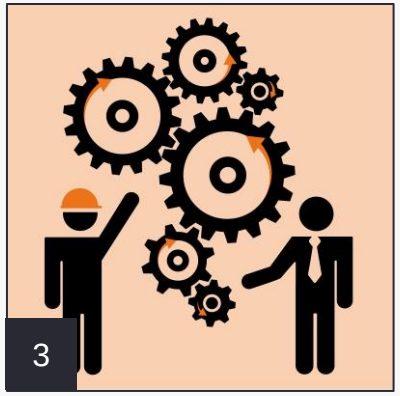


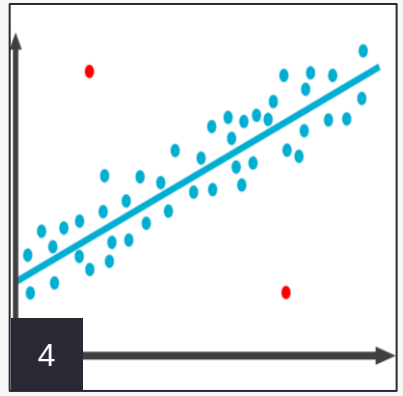


EDA Approach

1. Data Reading and Understanding
2. Missing values & Null Values check
3. Standardization
4. Outlier Identification
5. Data Analysis
6. Drawing Conclusions

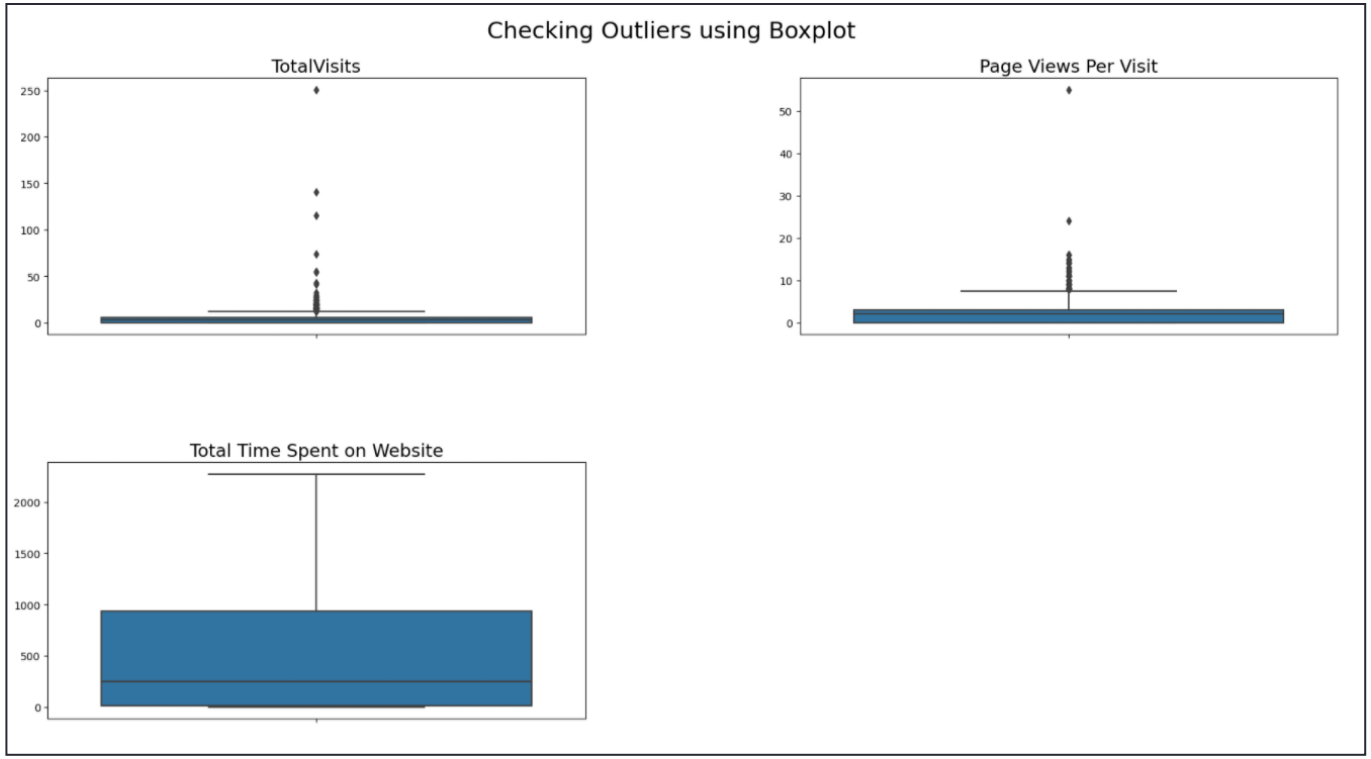




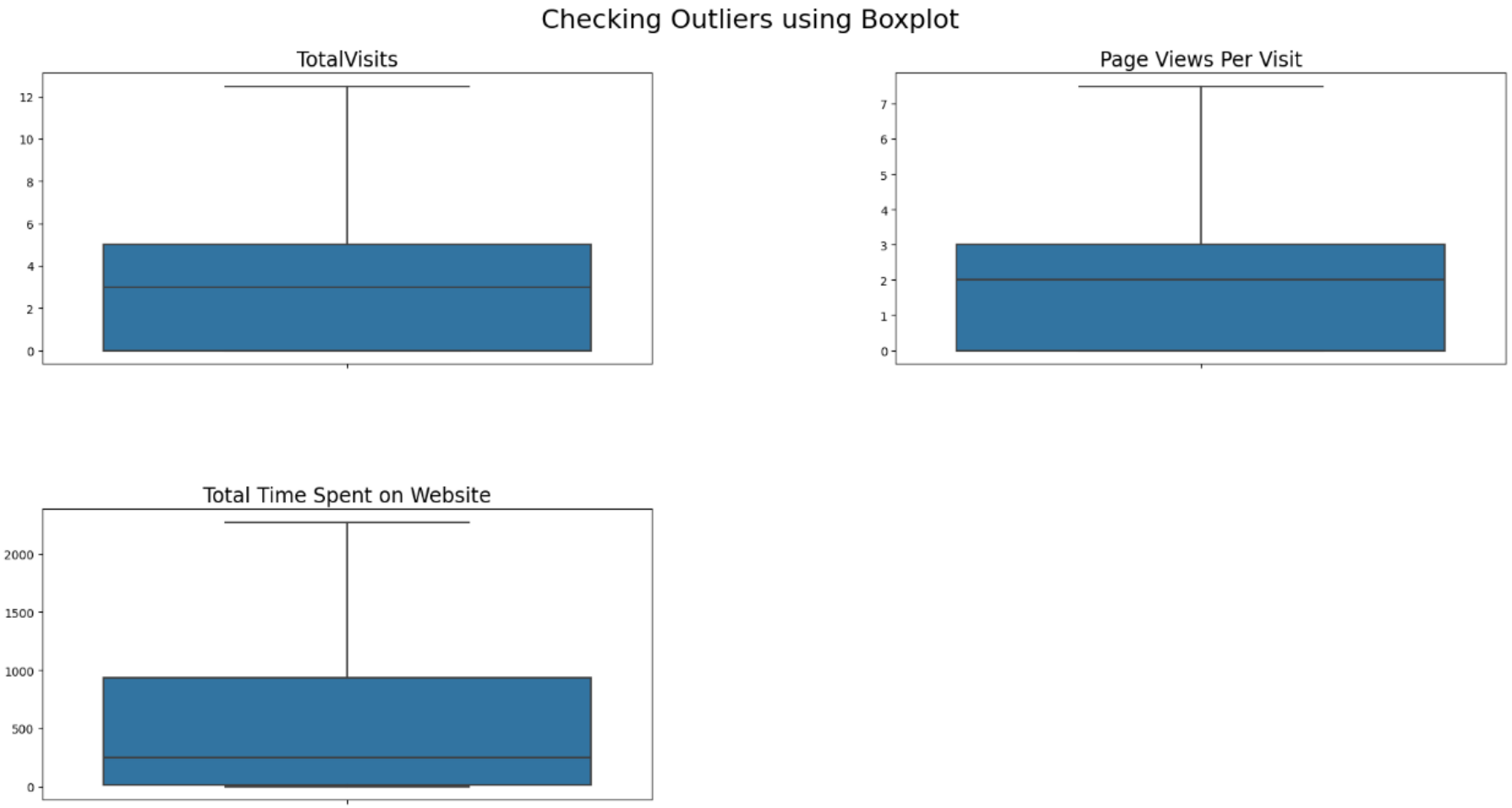


Outlier Analysis

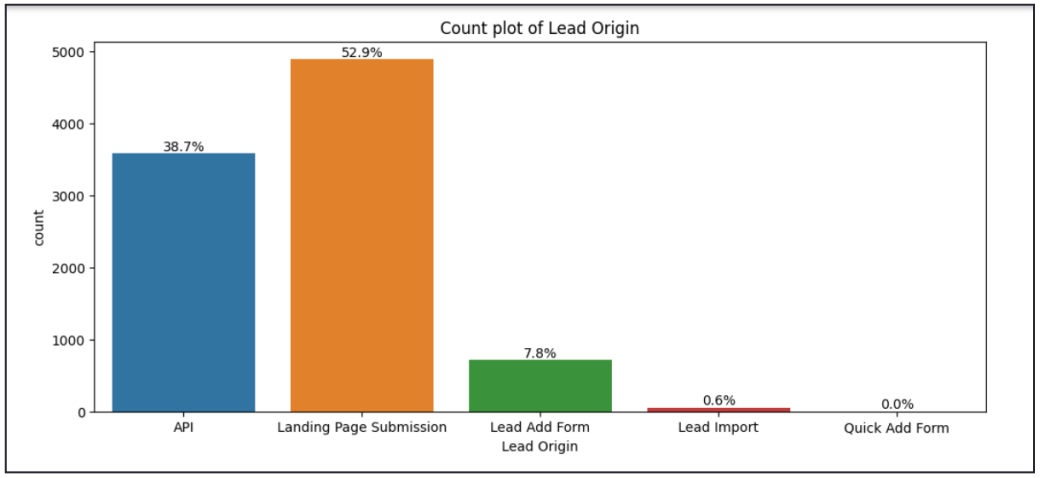
- The attribute “**totalvisits**” and “**Page views per visit**” contains few outliers in. the dataset , since these are quite less we can proceed to remove these outliers

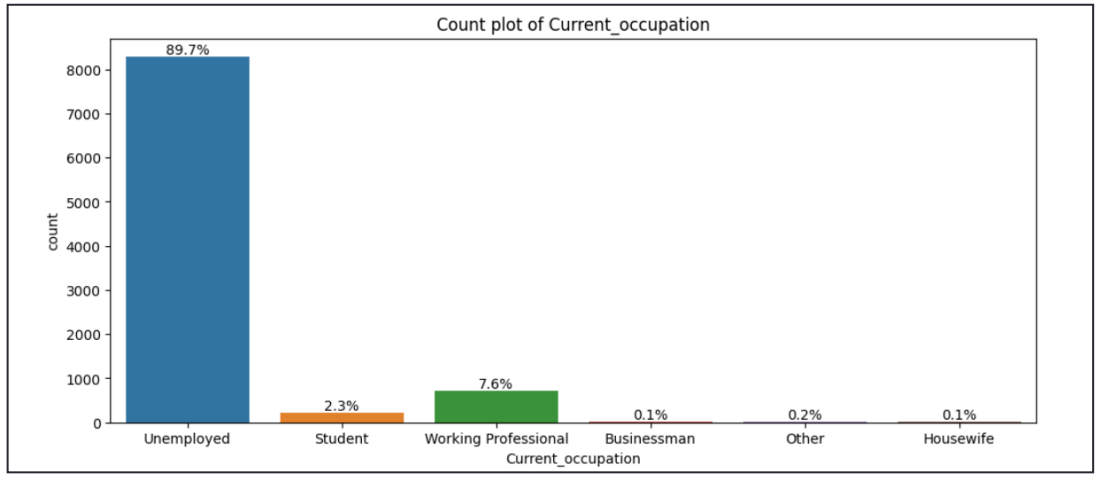


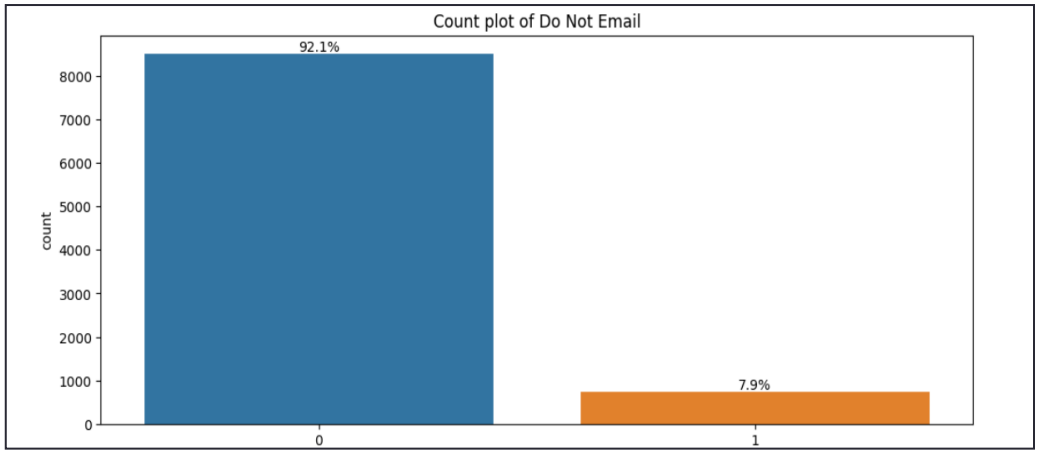
After Outlier Treatment

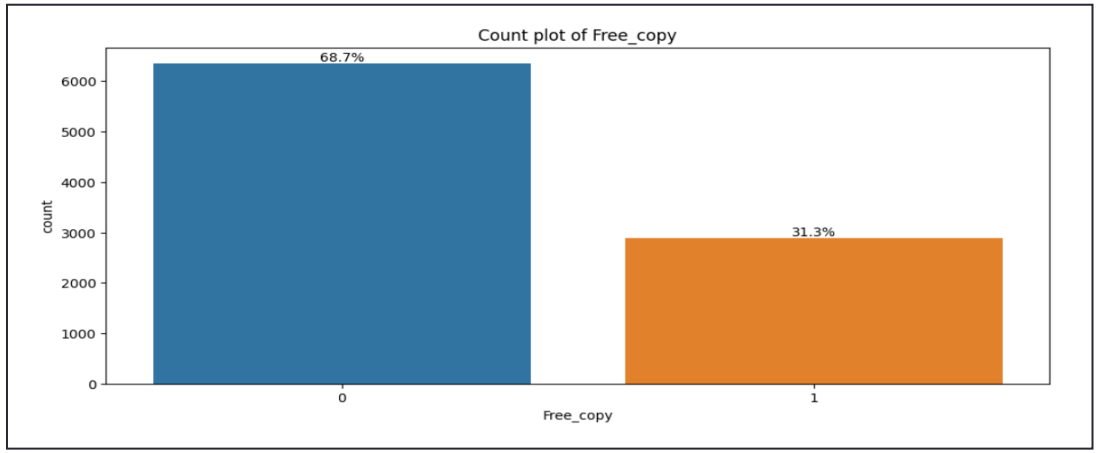


Univariate Analysis of Categorical Variables









Checking & Dropping Skewed Category Columns

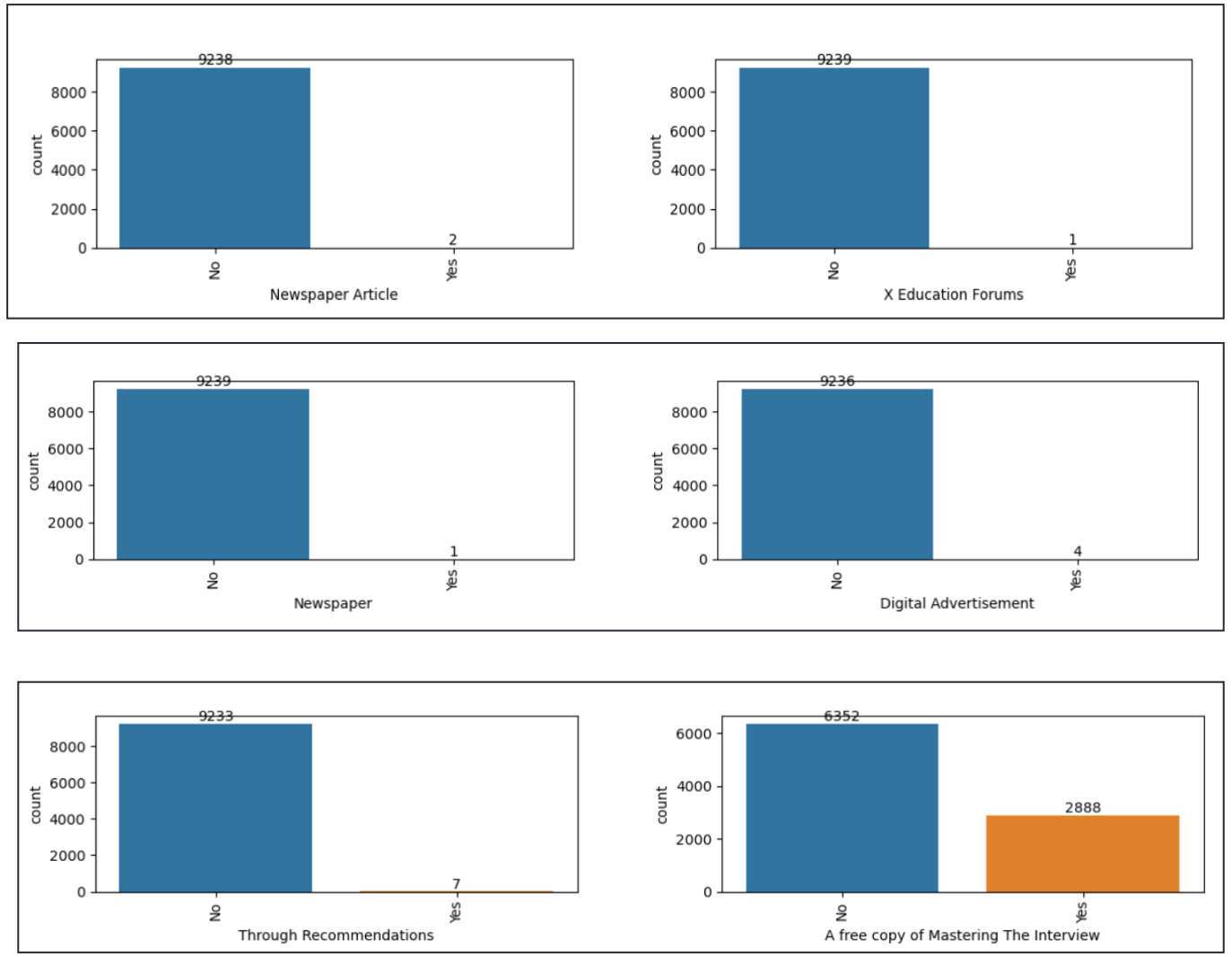
- **Observation:**

Columns with highly skewed data:

- 'Do Not Call',
- 'Search',
- 'Newspaper Article',
- 'X Education Forums',
- 'Newspaper',
- 'Digital Advertisement',
- 'Through Recommendations'.

- **Action:**
- These columns will be dropped. Reason: Skewed variables can affect the performance of logistic regression models, as they can get biased or they can give inaccurate parameter estimates.

Plot of skewed data

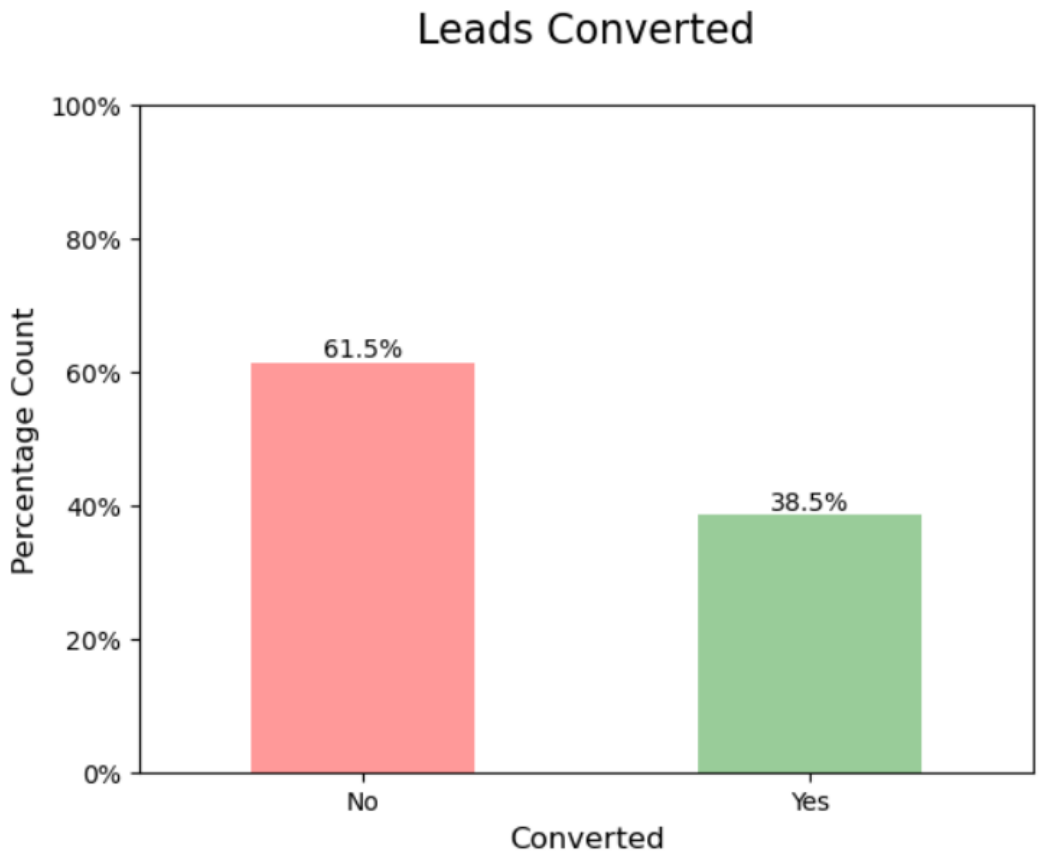


Check for Data Imbalance

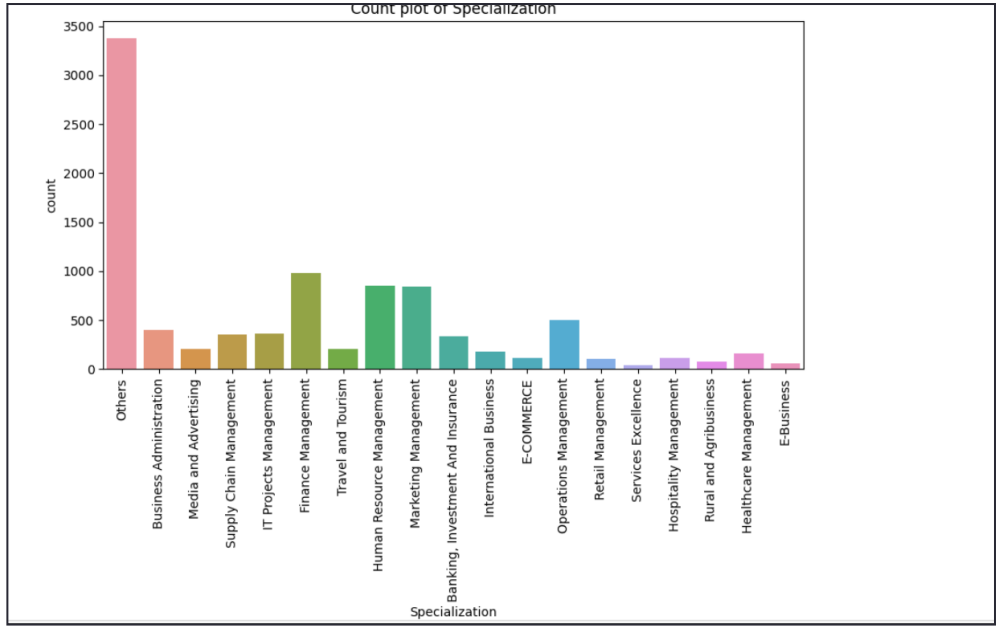
Observation:

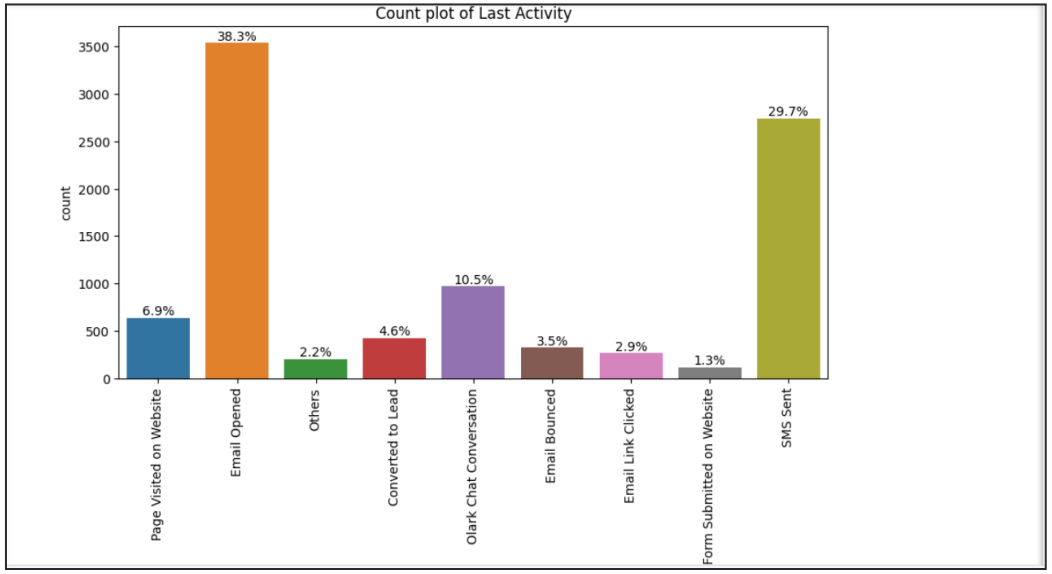
- 38.5% of the people have converted to leads.(M
- 61.5% of the people didnt convert to leads. (Ma

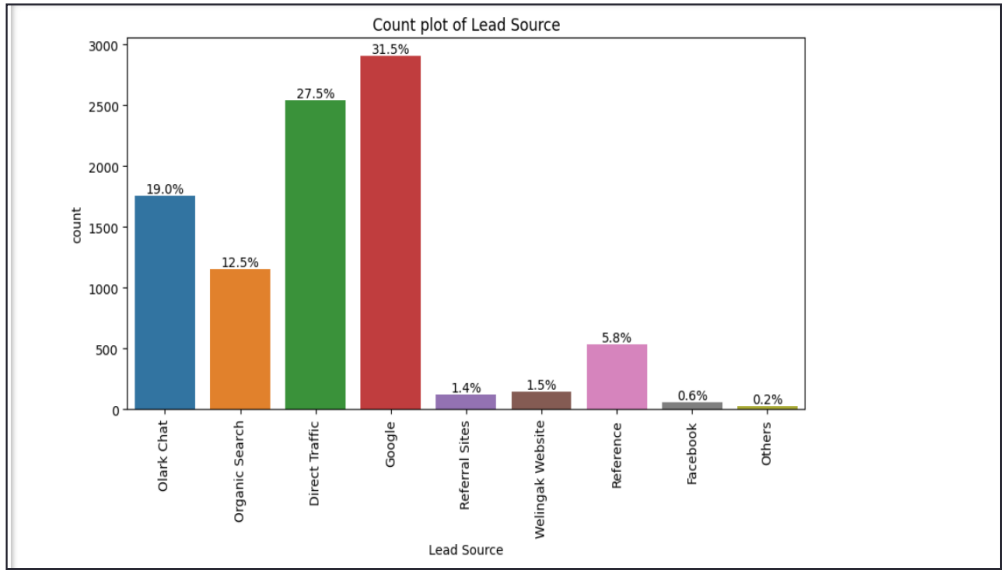
- Data Imbalance Ratio : 1.59 : 1



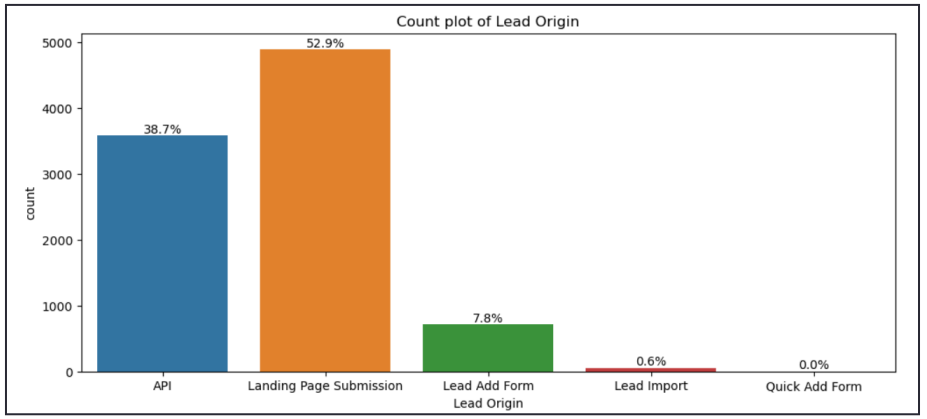
Bar-plot of Numerical Columns

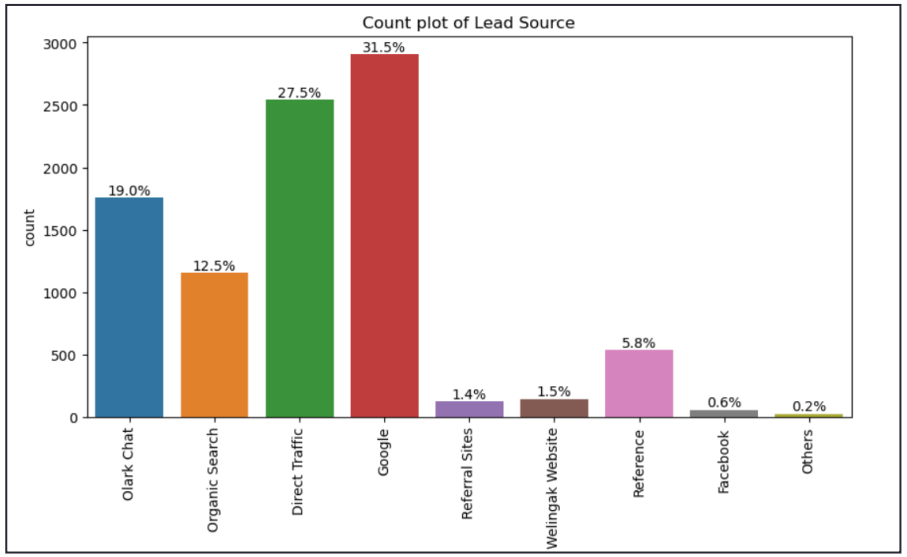


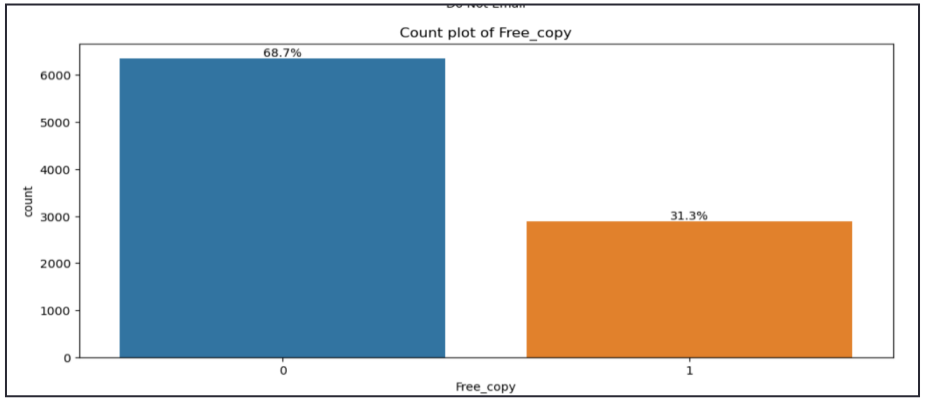


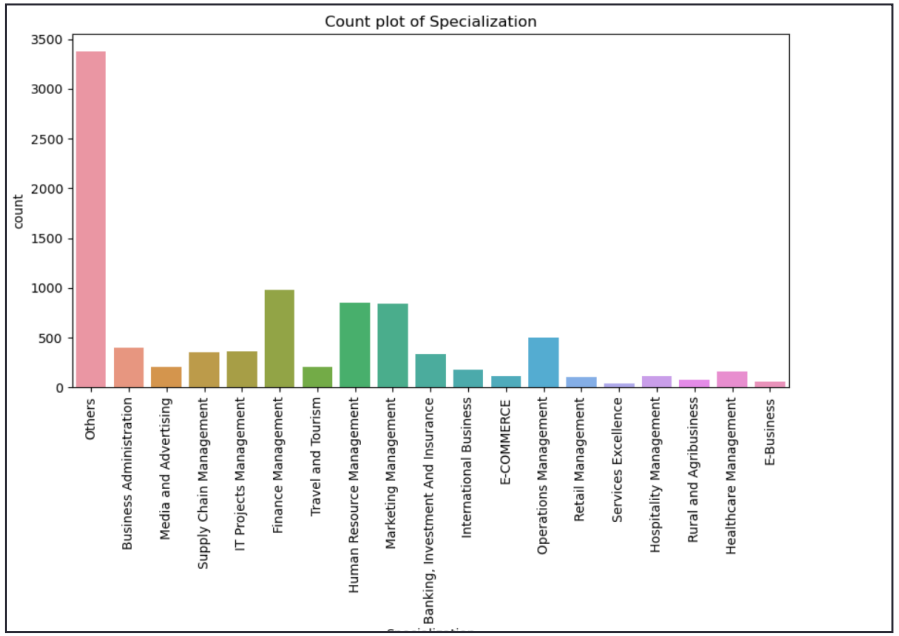


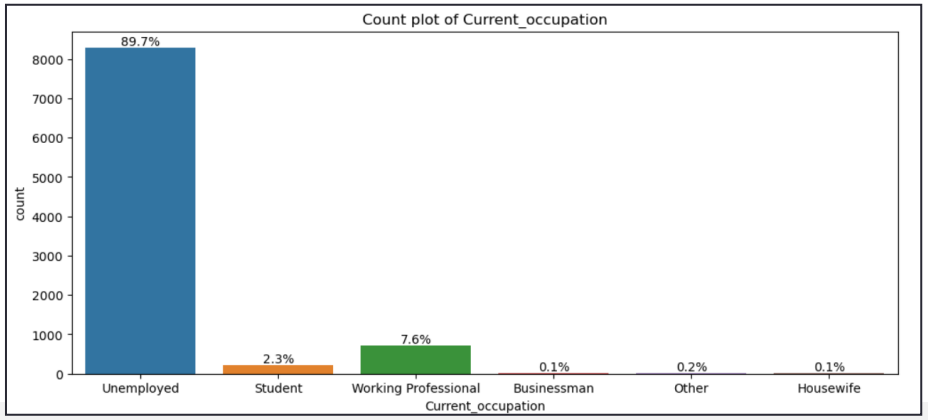
Univariate Analysis











Observations

List of features from variables which are present in majority (Converted and Not Converted included)

Lead Origin: "Landing Page Submission" identified 53% customers, "API" identified 39%.

Current_occupation: It has 90% of the customers as Unemployed

Do Not Email: 92% of the people has opted that they don’t want to be emailed about the course.

Lead Source: 58% Lead source is from Google & Direct Traffic combined

Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

Action:

These insights will be useful in further Bivariate Analysis.

Bivariate Analysis

Lead Origin: Around 52% of all leads originated from 'Landing Page Submission', with 36% lear conversion rate (LCR).The "API" identified approximately 39% of customers with 31% LCR.

Current_occupation: Around 90% of the customers are 'Unemployed' with 34% LCR. While 'Working Professional' contribute only 7.6% of total customers with almost 92% LCR.

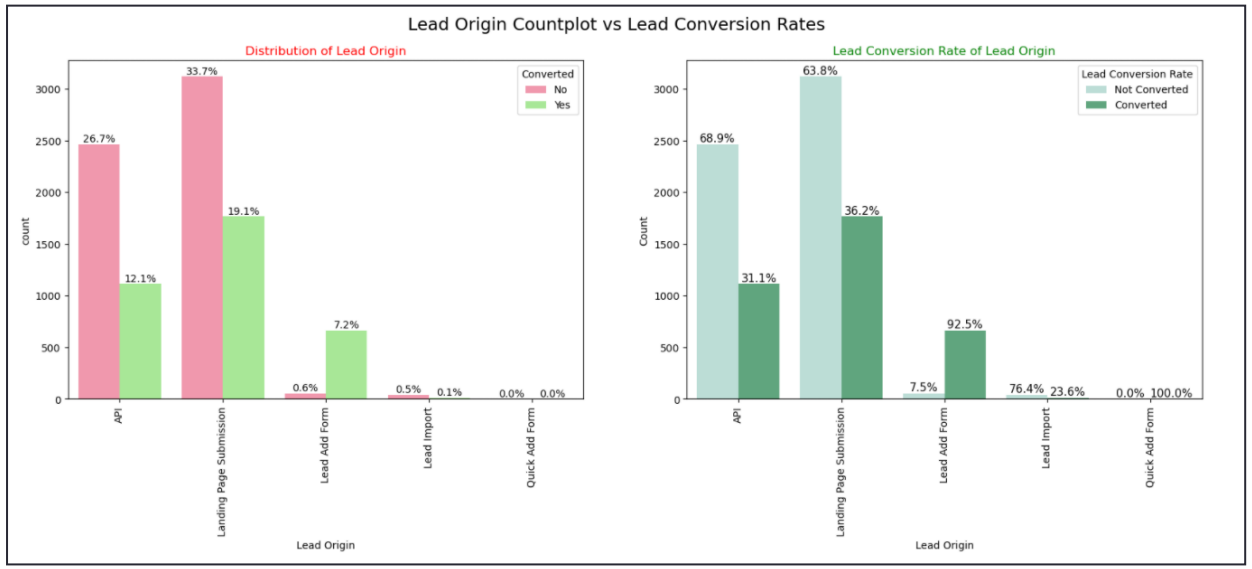
Do Not Email: 92% of the people has opted that they dont want to be emailed about the course.

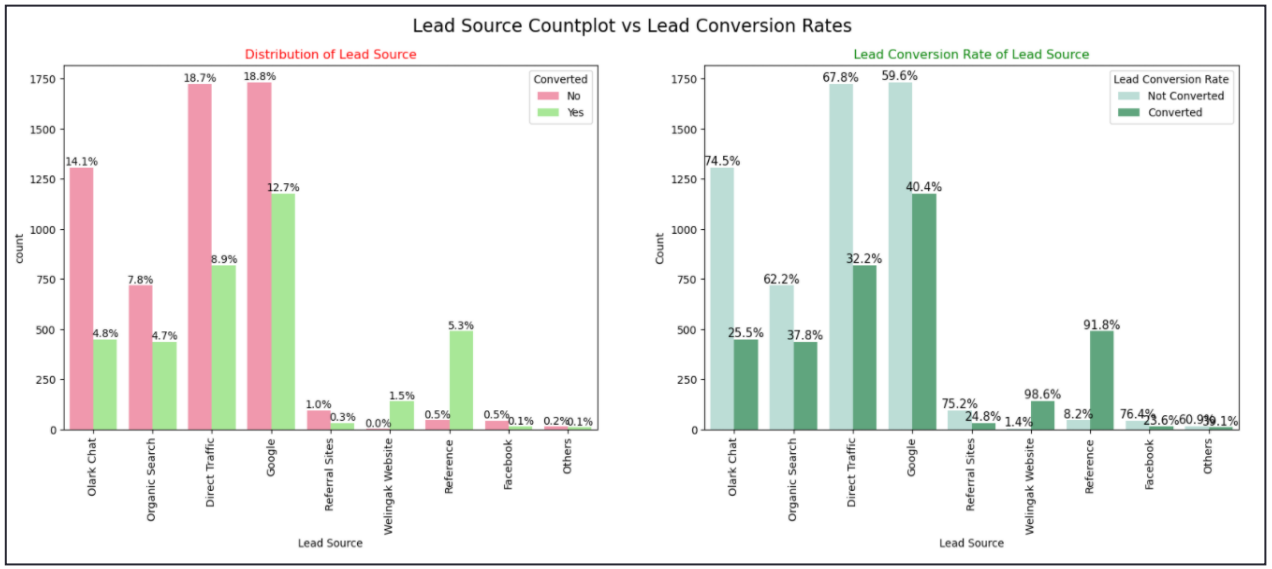
Lead Source: 'Google' has LCR of 40% out of 31% customers, 'Direct Traffic' contributes 32% LCR with 27% customers (lower than 'Google'), 'Organic Search' also gives 37.8% of LCR but the contribution is by only 12.5% of customers, 'Reference' has 91% LCR but there are only around 6% of customers through this Lead Source.

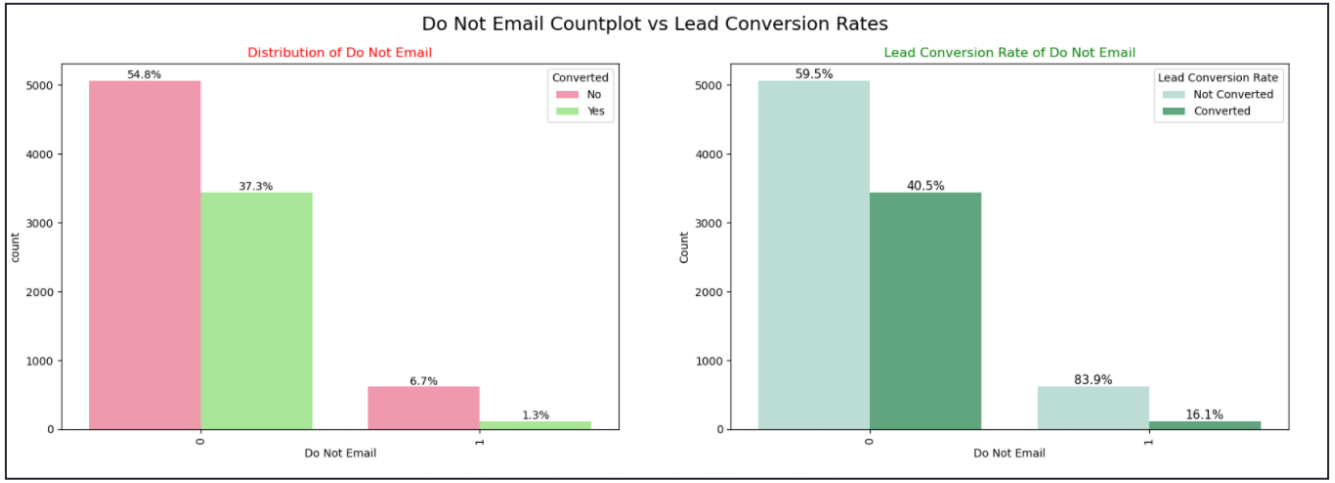
Last Activity: 'SMS Sent' has high lead conversion rate of 63% with 30% contribution from last activities, 'Email Opened' activity contributed 38% of last activities performed by the customers with 37% LCR.

Specialization: Marketing Management,HR Management,Finance Management shows good contribution.
(Abbreviation:**LCR** as **Lead Conversion Rate**)

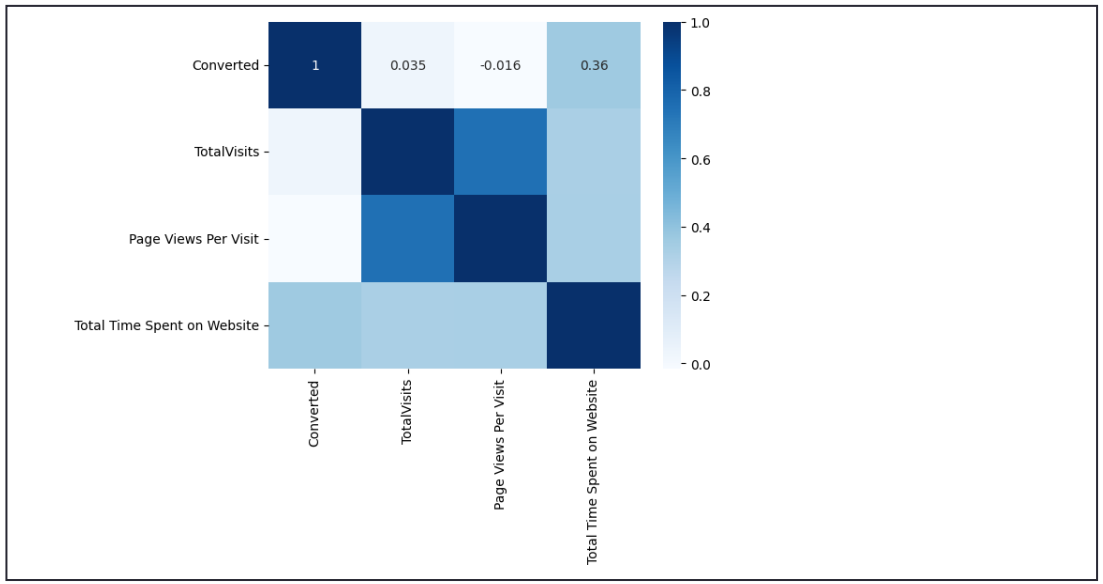
Plots Depicting various analytical outputs

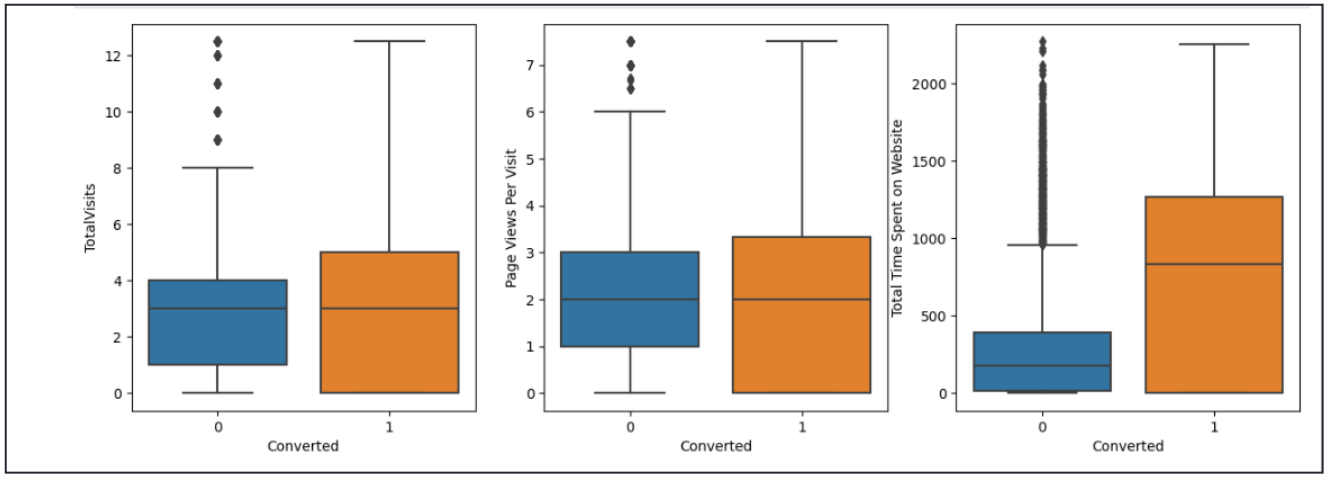






Continued…





- These can be validated through the above plots, clearly seen **Total visits** has a high correlation with **Page views per visit**
- And all the successful leads had **Totalvisits, Page views per visit, Total time spent on Website** all higher.

Dummy Variables and Feature Scaling

Dummy Variables were created for the following Columns:-

- **Lead Origin**
- **Lead Source**
- **Last Activity**
- **Specialization**
- **Current_occupation**

For Scaling of the columns we implemented StandardScaler.

Scaling & feature removal

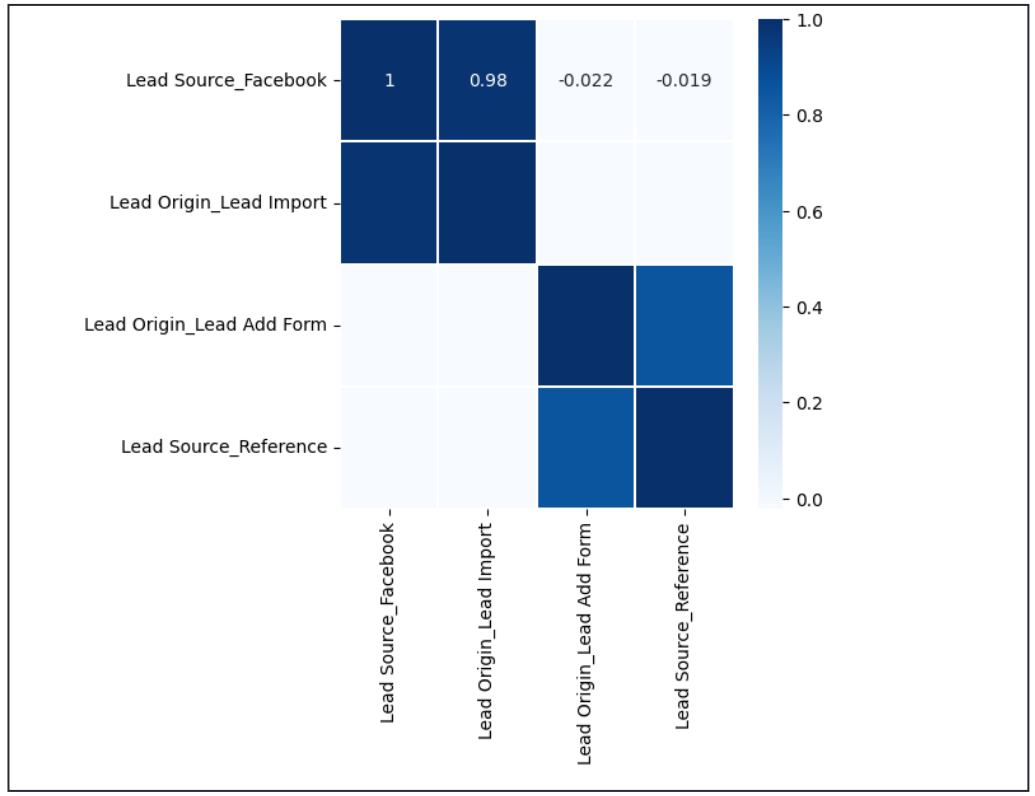
The following variables “**Lead source_facebook**”

“**Lead origin _lead Add form**” “**Lead Source_Reference**” has a higher positive correlation with each other”.

These variables are very highly correlated with each other (0.98 and 0.85), it is better that we drop one of these variables from each pair as they won’t add much value to the model.

Action:

Dropping 'Lead Origin_Lead Import' and 'Lead Origin_Lead Add Form'.



Model Building

Our approach for Model building:-

- Logistic Regression Model for predicting categorical v
- Feature Selection Using RFE (Coarse tuning)
- Manual fine-tuning using p-values and VIFs

Top 15 features chosen by Recursive Feature eliminatic

[256]:

	features	Feature Chosen	Ranking
22	Last Activity_SMS Sent	True	1
41	Current_occupation_Housewife	True	1
35	Specialization_Others	True	1
28	Specialization_Hospitality Management	True	1
20	Last Activity_Others	True	1
19	Last Activity_Olark Chat Conversation	True	1
17	Last Activity_Email Opened	True	1
14	Lead Source_Welingak Website	True	1
12	Lead Source_Reference	True	1
11	Lead Source_Others	True	1
45	Current_occupation_Working Professional	True	1
9	Lead Source_Olark Chat	True	1
7	Lead Source_Facebook	True	1
5	Lead Origin_Landing Page Submission	True	1
2	Total Time Spent on Website	True	1

Model 1 Results and Observations

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2732.8
Date:	Sat, 13 Jan 2024	Deviance:	5465.5
Time:	19:22:38	Pearson chi2:	8.09e+03
No. Iterations:	21	Pseudo R-squ. (CS):	0.3839
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0333	0.144	-7.155	0.000	-1.316	-0.750
Total Time Spent on Website	1.0505	0.039	27.169	0.000	0.975	1.126
Lead Origin_Landing Page Submission	-1.2721	0.126	-10.059	0.000	-1.520	-1.024
Lead Source_Facebook	-0.6961	0.529	-1.316	0.188	-1.733	0.340
Lead Source_Olark Chat	0.9001	0.119	7.585	0.000	0.668	1.133
Lead Source_Others	0.9807	0.512	1.915	0.056	-0.023	1.985
Lead Source_Reference	2.8977	0.216	13.434	0.000	2.475	3.320
Lead Source_Welingak Website	5.3802	0.729	7.384	0.000	3.952	6.808
Last Activity_Email Opened	0.9506	0.105	9.061	0.000	0.745	1.156
Last Activity_Olark Chat Conversation	-0.5534	0.187	-2.956	0.003	-0.920	-0.186
Last Activity_Others	1.2580	0.238	5.276	0.000	0.791	1.725
Last Activity_SMS Sent	2.0688	0.108	19.188	0.000	1.857	2.280
Specialization_Hospitality Management	-1.0720	0.324	-3.310	0.001	-1.707	-0.437
Specialization_Others	-1.1937	0.121	-9.841	0.000	-1.431	-0.956
Current_occupation_Housewife	23.0222	1.33e+04	0.002	0.999	-2.6e+04	2.6e+04
Current_occupation_Working Professional	2.6855	0.190	14.104	0.000	2.312	3.059

"**Current_occupation_Housewife**" column will be removed from model due to high p-value of 0.999.

Model 2 Results and Observations

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6453
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2740.3
Date:	Sat, 13 Jan 2024	Deviance:	5480.7
Time:	19:22:40	Pearson chi2:	8.12e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3825
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0251	0.144	-7.111	0.000	-1.308	-0.743
Total Time Spent on Website	1.0494	0.039	27.177	0.000	0.974	1.125
Lead Origin_Landing Page Submission	-1.2674	0.126	-10.030	0.000	-1.515	-1.020
Lead Source_Facebook	-0.6969	0.529	-1.318	0.187	-1.733	0.339
Lead Source_Olark Chat	0.8991	0.119	7.580	0.000	0.667	1.132
Lead Source_Others	0.9739	0.512	1.902	0.057	-0.030	1.977
Lead Source_Reference	2.9171	0.215	13.538	0.000	2.495	3.339
Lead Source_Welingak Website	5.3791	0.729	7.384	0.000	3.951	6.807
Last Activity_Email Opened	0.9490	0.105	9.077	0.000	0.744	1.154
Last Activity_Olark Chat Conversation	-0.5583	0.187	-2.985	0.003	-0.925	-0.192
Last Activity_Others	1.2482	0.238	5.238	0.000	0.781	1.715
Last Activity_SMS Sent	2.0588	0.108	19.151	0.000	1.848	2.270
Specialization_Hospitality Management	-1.0795	0.324	-3.334	0.001	-1.714	-0.445
Specialization_Others	-1.1978	0.121	-9.881	0.000	-1.435	-0.960
Current_occupation_Working Professional	2.6773	0.190	14.068	0.000	2.304	3.050

"**Lead Source_Facebook**" column will be removed from model (high p-value=0.187).

Final Model Evaluation Results

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2743.1
Date:	Sat, 13 Jan 2024	Deviance:	5486.1
Time:	19:22:45	Pearson chi2:	8.11e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3819
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0236	0.143	-7.145	0.000	-1.304	-0.743
Total Time Spent on Website	1.0498	0.039	27.234	0.000	0.974	1.125
Lead Origin_Landing Page Submission	-1.2590	0.125	-10.037	0.000	-1.505	-1.013
Lead Source_Olark Chat	0.9072	0.118	7.701	0.000	0.676	1.138
Lead Source_Reference	2.9253	0.215	13.615	0.000	2.504	3.346
Lead Source_Welingak Website	5.3887	0.728	7.399	0.000	3.961	6.816
Last Activity_Email Opened	0.9421	0.104	9.022	0.000	0.737	1.147
Last Activity_Olark Chat Conversation	-0.5556	0.187	-2.974	0.003	-0.922	-0.189
Last Activity_Others	1.2531	0.238	5.259	0.000	0.786	1.720
Last Activity_SMS Sent	2.0519	0.107	19.106	0.000	1.841	2.262
Specialization_Hospitality Management	-1.0944	0.323	-3.391	0.001	-1.727	-0.462
Specialization_Others	-1.2033	0.121	-9.950	0.000	-1.440	-0.966
Current_occupation_Working Professional	2.6697	0.190	14.034	0.000	2.297	3.042

	Features	VIF
0	Specialization_Others	2.47
1	Lead Origin_Landing Page Submission	2.45
2	Last Activity_Email Opened	2.36
3	Last Activity_SMS Sent	2.20
4	Lead Source_Olark Chat	2.14
5	Last Activity_Olark Chat Conversation	1.72
6	Lead Source_Reference	1.31
7	Total Time Spent on Website	1.24
8	Current_occupation_Working Professional	1.21
9	Lead Source_Welingak Website	1.08
10	Last Activity_Others	1.08
11	Specialization_Hospitality Management	1.02

Model 4 is stable and has significant p-values within the threshold (p-values < 0.05)

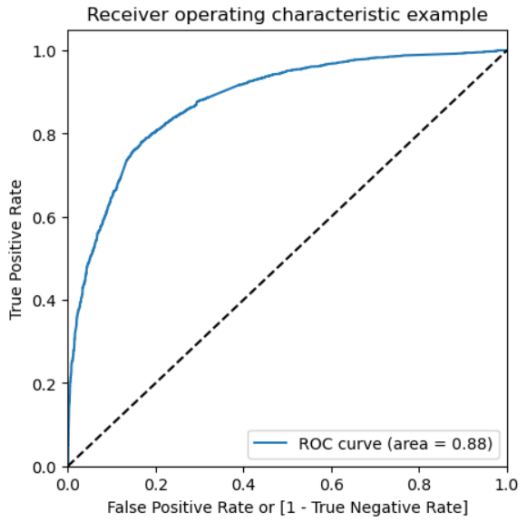
No variable to drop as they all have good VIF values (less than 5).

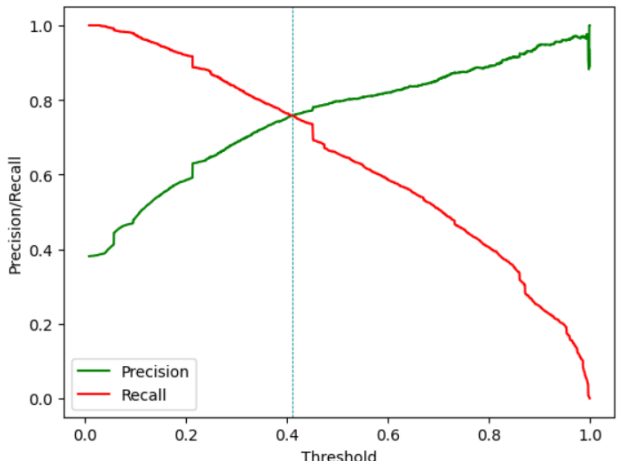
P-values for all variables is less than 0.05.

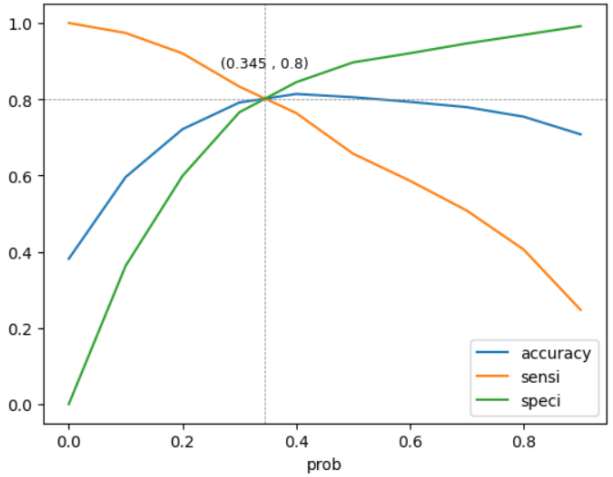
Model Evaluation results on Train Set

Observation:

- Area under ROC curve is 0.88 out of 1 which indicates a good predictive m
- Accuracy,sensitivity and specificity for various probability cutoffs
- 0.345 seems to be the Optimal cutoff point for probability threshold.
- The intersection point of the Precision/Recall curve is the threshold value the model achieves a balance between precision and recall, which can be to optimise the performance of the model based on business requiremen





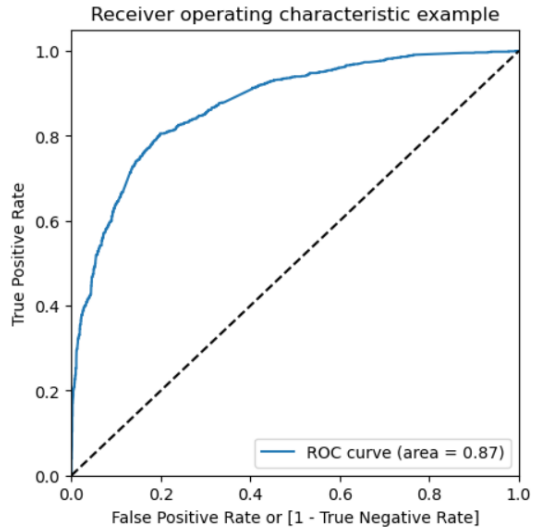


	prob	accuracy	sensi	speci
0.0	0.0	0.381262	1.000000	0.000000
0.1	0.1	0.595702	0.973642	0.362819
0.2	0.2	0.721243	0.920114	0.598701
0.3	0.3	0.791280	0.832928	0.765617
0.4	0.4	0.813698	0.763585	0.844578
0.5	0.5	0.805195	0.656934	0.896552
0.6	0.6	0.792981	0.585969	0.920540
0.7	0.7	0.779066	0.507705	0.946277
0.8	0.8	0.754020	0.405515	0.968766
0.9	0.9	0.707792	0.247364	0.991504

Model Evaluation results on Test Set

Observation

- Area under ROC curve is 0.87 out of 1 which indicates a good predictive m



For Test set

Accuracy : 80.34%

Sensitivity : 79.82% ≈ 80%

Specificity : 80.68%

- The customers with a higher lead score have a higher conversion chance compared to the ones with lower lead sco

	Prospect ID	Converted	Converted_Prob	final_predicted	Lead_Score
0	4269	1	0.697934	1	70
1	2376	1	0.860665	1	86
2	7766	1	0.889241	1	89
3	9199	0	0.057065	0	6
4	4359	1	0.871510	1	87

Conclusions

Model parameters

- The final Logistic Regression Model has 12 features

Top 3 features that contributing 'positively' to predicting good leads are:-

- Lead Source_Welingak Website
- Lead Source_Reference
- Current_occupation_Working Professional

Optimal cutoff probability point = 0.345. Converted probability greater than 0.345 will be predicted as Converted lead & probability smaller than 0.345 will be predicted as not Converted lead.

Recommendations

To increase our Lead Conversion Rates:-

Focus on features with positive coefficients for targeted marketing strategies.

Develop strategies to attract high-quality leads from top-performing lead sources.

Engage working professionals with tailored messaging.

Optimize communication channels based on lead engagement impact.

More budget/spend can be done on Welingak Website in terms of advertising, etc.

Incentives/discounts for providing reference that convert to lead, encourage providing more references.

Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement:-

Analyze negative coefficients in specialization offerings.

Review landing page submission process for areas of improvement.

Thank You

Kashish ,Chandan,
Kanishka