

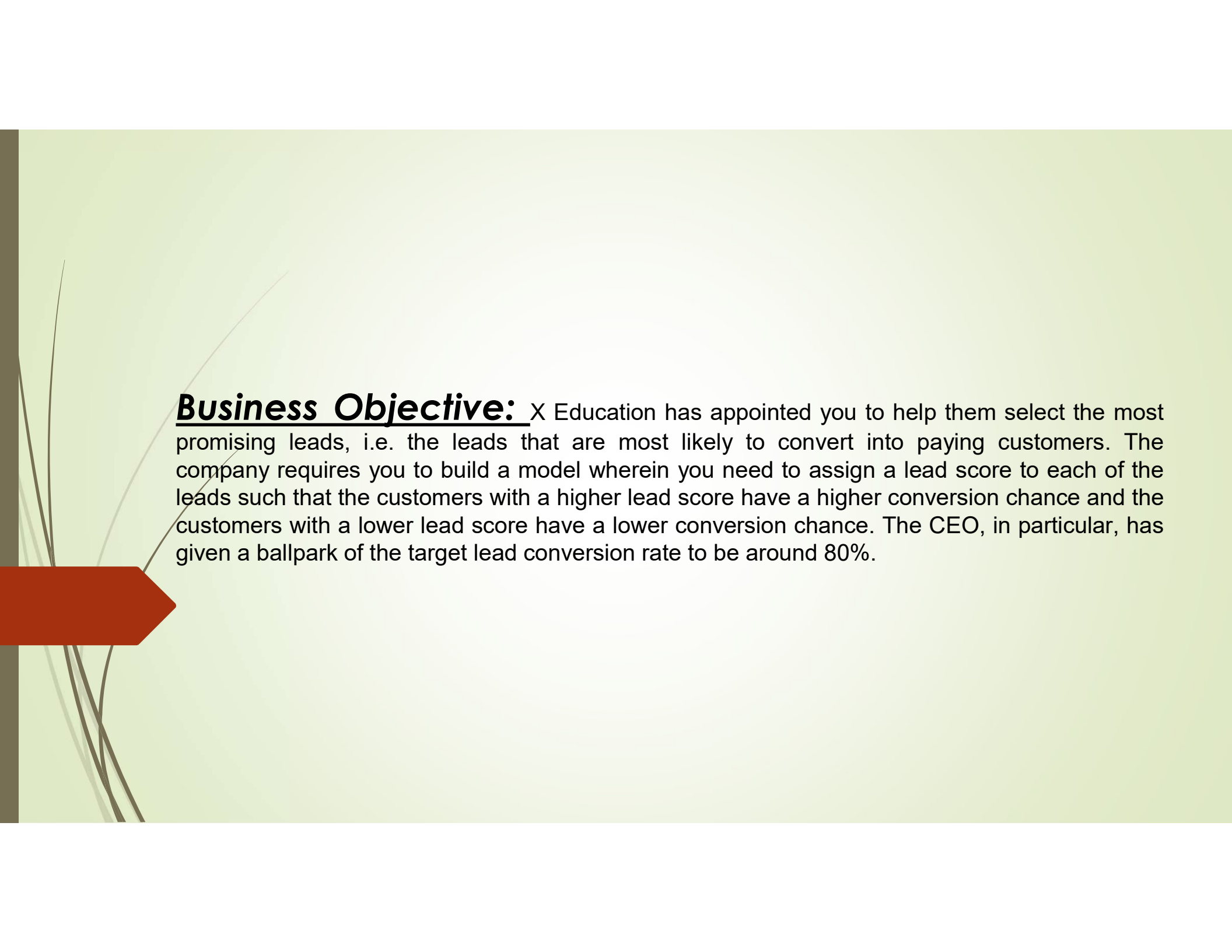
# **REPORT ON LOGISTIC REGRESSION ASSIGNMENT – LEAD SCORING(Dated 16<sup>th</sup> Oct 2024)**

***Presented by***

**MR. VIBHUDENDRA S**

**upGrad & IIITB | Data Science Program - February 2024**

**Batch ID 5702**



**Business Objective:** X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

## Data Sources

- 'Leads.csv': contains all the information of the client on leads from the past
- '*Leads Data Dictionary.xlsx*' is data dictionary which describes the meaning of the variables.

# Business Problem Solving Methodology

**Problem solving approach methodology:** Data Science methodology, Machine Learning: Logistic Regression as this is a classification problem, Model evaluation, Insights & recommendations based on feature importance.

Steps followed are as under.

1. Importing the available data set
2. Data Understanding
3. Data Pre-processing
4. EDA
5. Data Preparation
6. Machine Learning/Model building
7. Model evaluation
8. Final model
9. Feature importance
10. Recommendations

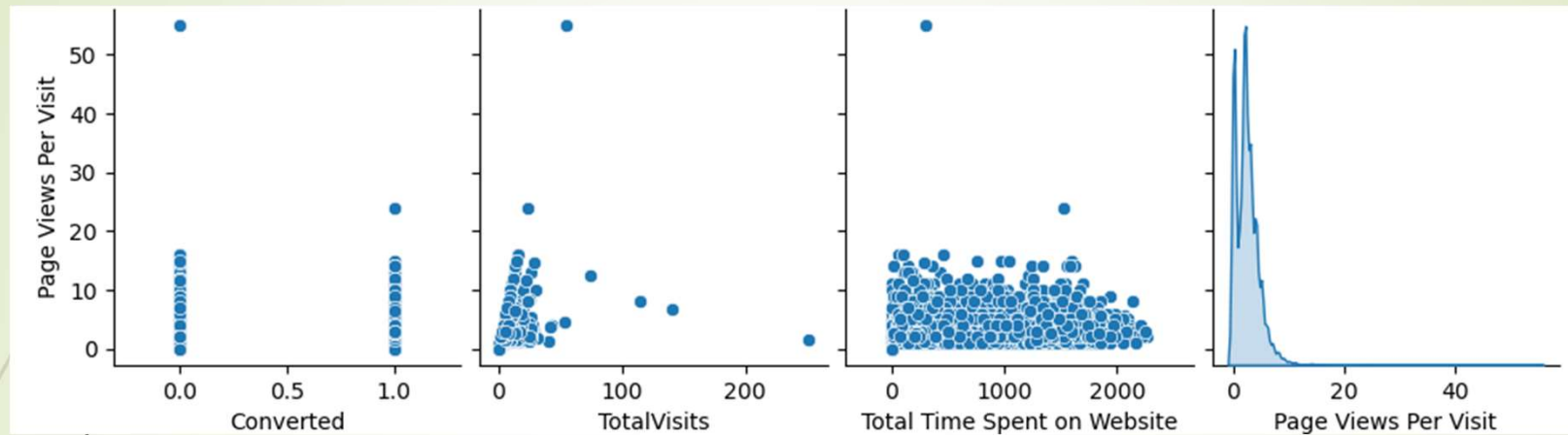
# Data Understanding & Pre-processing

This dataset has **9,240 records & 37 features**

There are **30 categorical features & 7 numerical features**

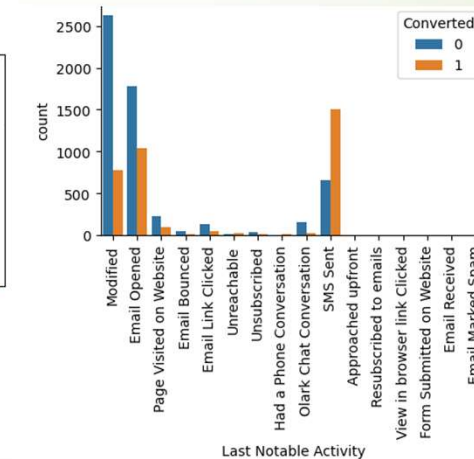
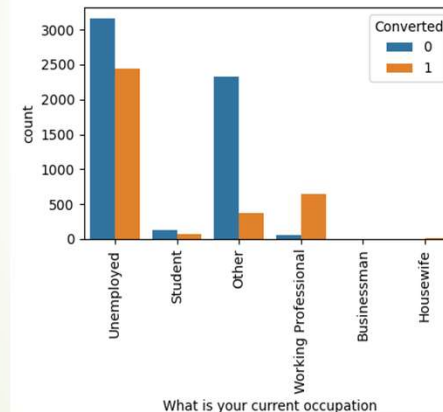
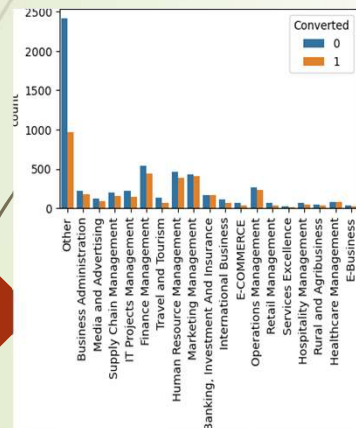
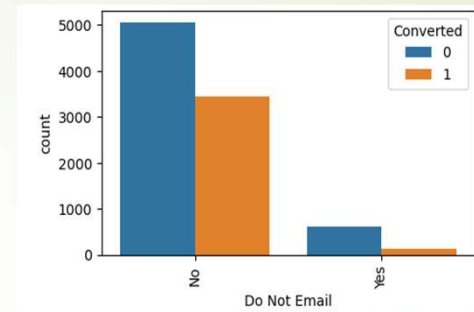
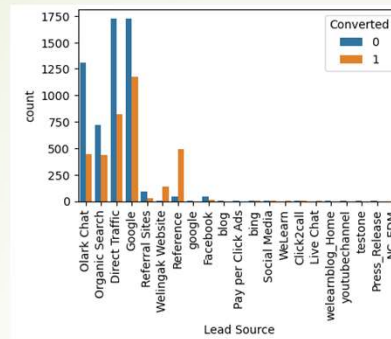
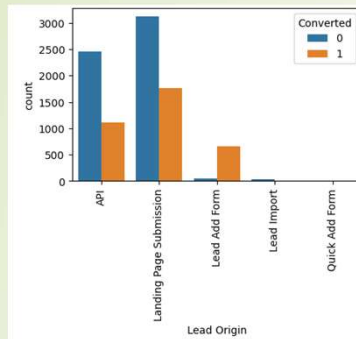
- **Checked & removed duplicate records**, but in this case there was no duplicate records found.
- There were few features which had 'select' as value which was made 'null' as applicable in this case.
- Null values handling: **7 features had > 40% null values** which we **dropped**. **10 more features** had **null values < 40%** which we **dropped few & imputed median & mode values** for rest of them. We **ended up** having **25 features** at this stage.
- **Checked for class imbalance: Dropped 13 features** which had highly imbalanced data.

# EDA

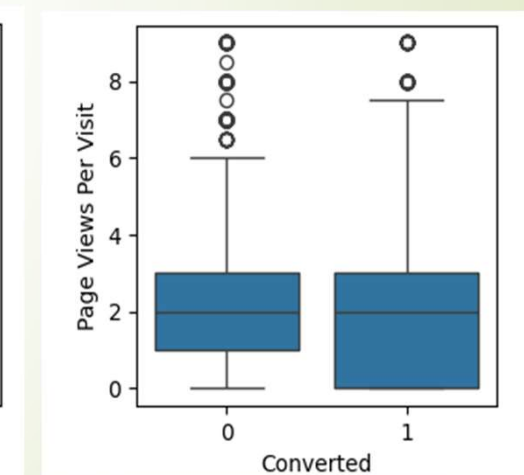
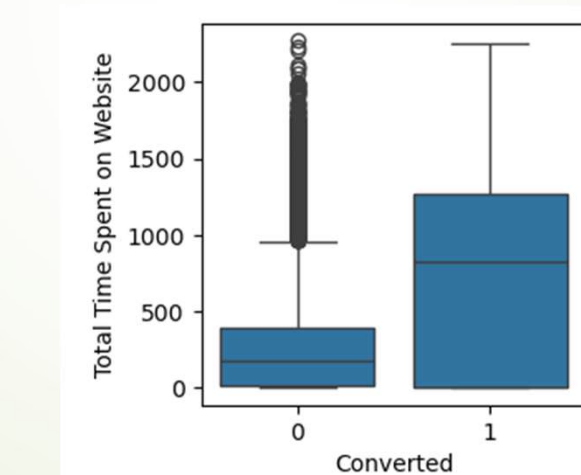
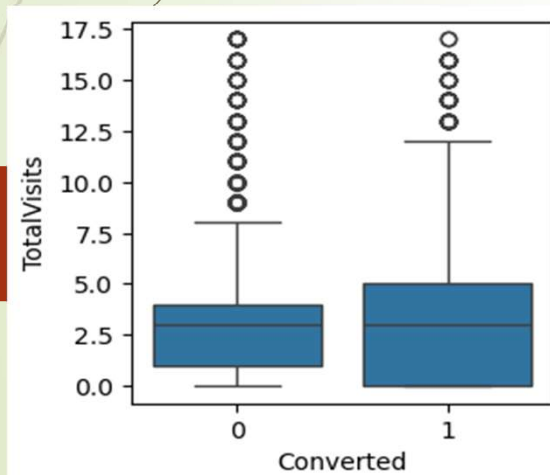
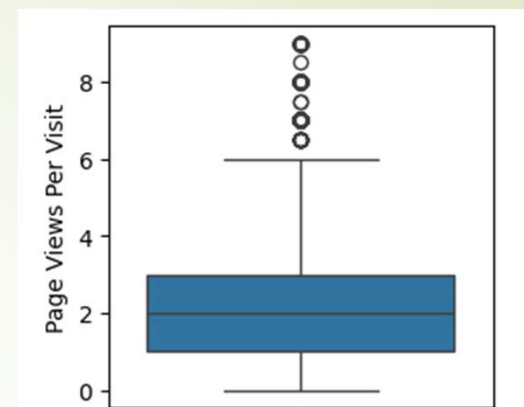
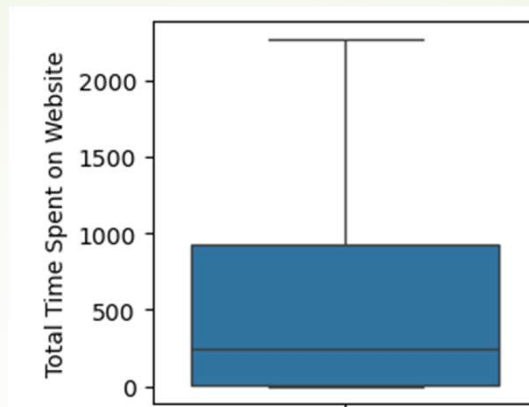
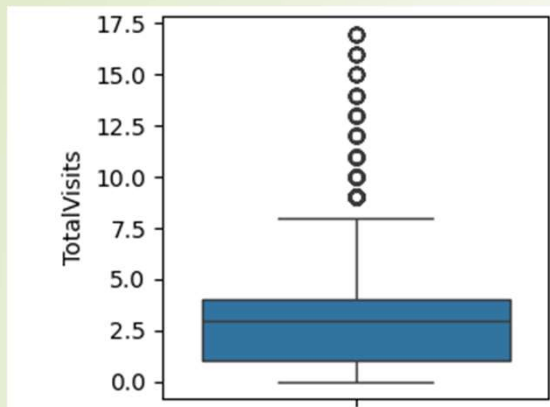


**Features 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit' have no linear correlation with target feature 'Converted'**

# EDA – COUNTPLOTS

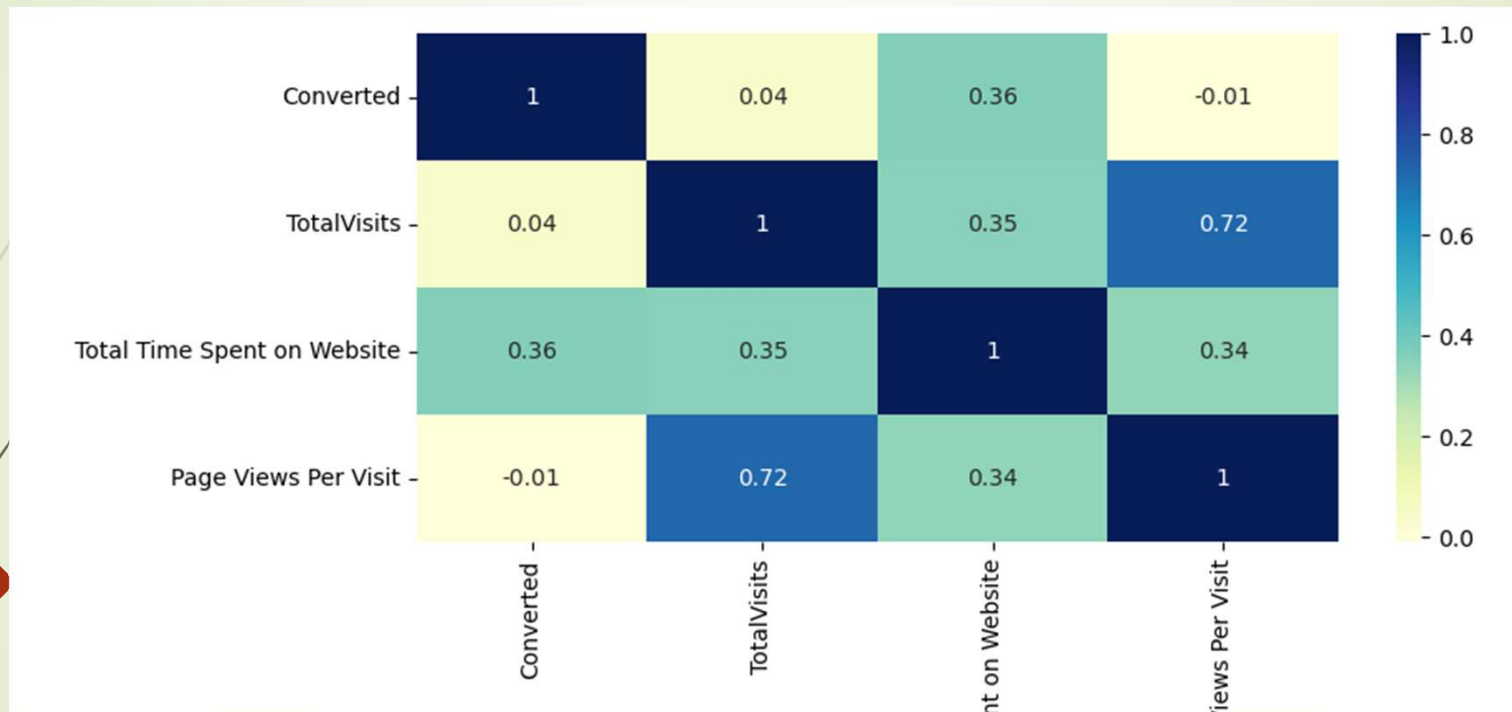


# EDA - BOXPLOTS





## EDA - HEATMAP



**'Page Views Per Visit' & 'Total Visits' have a high correlation of 0.72**

## EDA – Insights

- Lead Origins like 'API' & 'Landing Page Submission' have significant impact on lead conversions.
- Lead Sources like 'Google', 'Direct Traffic', 'Reference', 'Olark Chat' & 'Organic Search' have significant impact on lead conversions.
- Those who have opted for email have significant impact on lead conversions.
- If the Last Activity is either 'SMS sent' or 'Email opened', then they have significant impact on lead conversions.
- Those who have mentioned specialization as 'Others' have significant impact on lead conversions.
- Those who are 'Unemployed' have significant impact on lead conversions.
- Those who have not opted for 'A free copy of Mastering The Interview' have significant impact on lead conversions
- Those who have visited the website atleast 5 times have significant impact on lead conversion rate.
- Those who have spent total time on website of around 1200 units have significant impact on lead conversion rate.
- Those who have viewed Average number of pages on the website of atleast 3 have significant impact on lead conversion rate

## Outliers & its handling

- 'Total Visits' & 'Page Views Per Visit' have outliers on the upper range & required Outlier Treatment. From describe function previously, it was observed that 75th percentile value of 'Total Visits' is 5 & for 'Page Views Per Visit' is 3. Hence capped maximum values to 99th percentile.

## Data Preparation

- Converted 2 features binary values (Yes/No) to 1/0
- Converted 5 features 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization' & 'What is your current occupation' into numerical features using dummies
- We had totally all 11 as numerical features at this stage which was good enough for model building.
- Feature Scaling: on 3 features using Standard Scaler from sklearn
- We had almost 38% lead converted rate.
- 'Do Not Email' & 'Last Activity\_email bounced' have high correlation of 0.63.
- 'Lead Origin\_Lead Add Form' and 'Lead Source\_Referance' having higher correlation of 0.85.
- 'Lead Source\_Facebook' and 'Lead Origin\_Lead Import' having higher correlation of 0.98
- 'TotalVisits' and 'Page Views Per Visit' having correlation of 0.72

# Machine Learning/Model Building

- Used Logistic Regression from sklearn
- Feature selection using RFE(selected 15 important features)
- Used statsmodels to add constant & fit the model
- Used summary()/mainly p-values & VIF scores to tune the model
- Final model(Third model) had 13 features, all significant p-values & all VIF scores below 5.
- Using final model, predicted train dataset values.
- Found out Optimal cut-off point using accuracy, sensitivity & specificity scores as 0.35
- Tuned predicted values based on 0.35 threshold & added lead scores based on the same.
- The cut-off point of Precision Recall curve was around 0.45
- Plotted ROC curve & had AUC score of 0.89

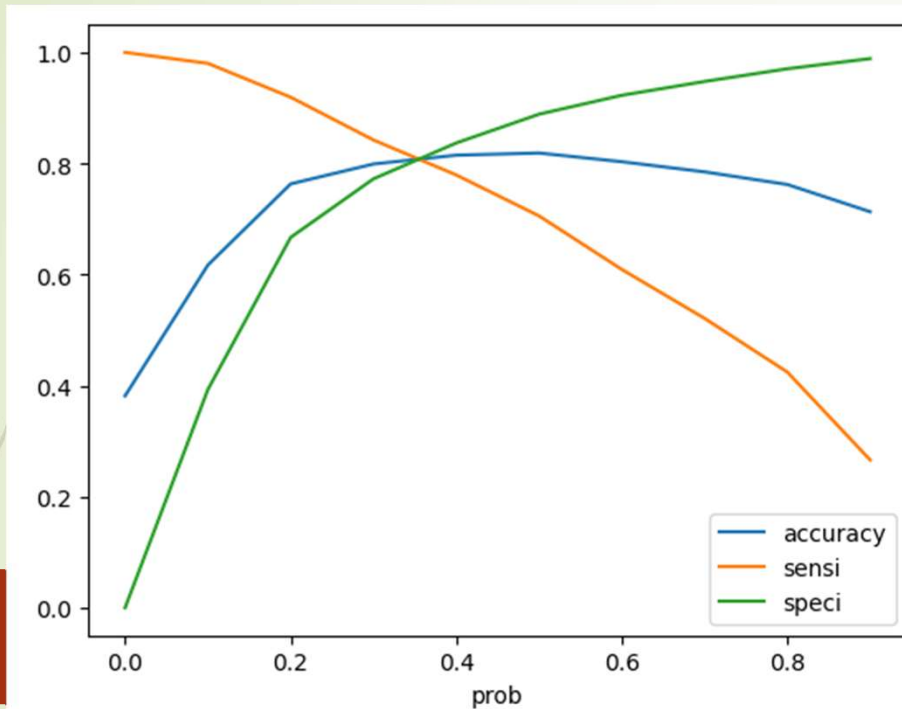
# Machine Learning/Model building(Contd..)

Generalized Linear Model Regression Results			
<b>Dep. Variable:</b>	Converted	<b>No. Observations:</b>	6363
<b>Model:</b>	GLM	<b>Df Residuals:</b>	6349
<b>Model Family:</b>	Binomial	<b>Df Model:</b>	13
<b>Link Function:</b>	Logit	<b>Scale:</b>	1.0000
<b>Method:</b>	IRLS	<b>Log-Likelihood:</b>	-2564.5
<b>Date:</b>	Wed, 16 Oct 2024	<b>Deviance:</b>	5129.0
<b>Time:</b>	23:32:36	<b>Pearson chi2:</b>	6.79e+03
<b>No. Iterations:</b>	7	<b>Pseudo R-squ. (CS):</b>	0.4076
<b>Covariance Type:</b>	nonrobust		

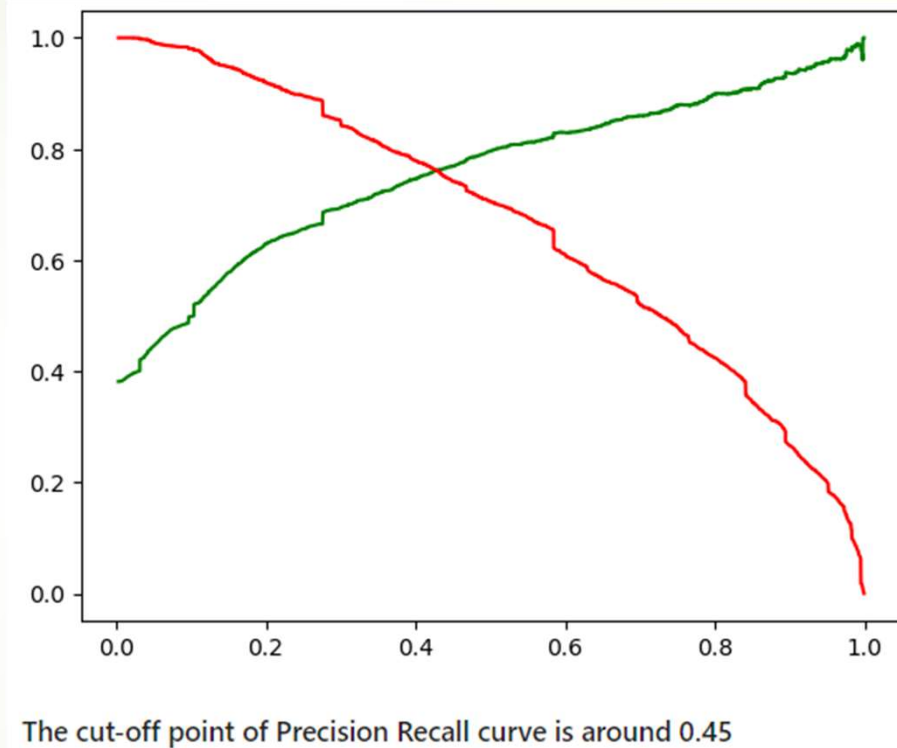
	Features	VIF
10	Specialization_Other	2.27
4	Lead Source_Olark Chat	2.06
2	Lead Origin_Landing Page Submission	1.72
11	What is your current occupation_Other	1.62
8	Last Activity_SMS Sent	1.54
3	Lead Origin_Lead Add Form	1.52
7	Last Activity_Olark Chat Conversation	1.49
5	Lead Source_Welingak Website	1.32
1	Total Time Spent on Website	1.25
0	Do Not Email	1.21
12	What is your current occupation_Working Profes...	1.20
9	Last Activity_Unsubscribed	1.10
6	Last Activity_Converted to Lead	1.09

	coef	std err	z	P> z	[0.025	0.975]
<b>const</b>	-0.2785	0.127	-2.195	0.028	-0.527	-0.030
<b>Do Not Email</b>	-1.6051	0.179	-8.962	0.000	-1.956	-1.254
<b>Total Time Spent on Website</b>	1.0675	0.040	26.445	0.000	0.988	1.147
<b>Lead Origin_Landing Page Submission</b>	-0.8760	0.130	-6.753	0.000	-1.130	-0.622
<b>Lead Origin_Lead Add Form</b>	2.8909	0.211	13.726	0.000	2.478	3.304
<b>Lead Source_Olark Chat</b>	1.0952	0.125	8.787	0.000	0.851	1.339
<b>Lead Source_Welingak Website</b>	3.1667	1.029	3.078	0.002	1.150	5.183
<b>Last Activity_Converted to Lead</b>	-1.1395	0.210	-5.422	0.000	-1.551	-0.728
<b>Last Activity_Olark Chat Conversation</b>	-1.2742	0.167	-7.635	0.000	-1.601	-0.947
<b>Last Activity_SMS Sent</b>	1.3083	0.076	17.197	0.000	1.159	1.457
<b>Last Activity_Unsubscribed</b>	1.3908	0.452	3.080	0.002	0.506	2.276
<b>Specialization_Other</b>	-0.8366	0.124	-6.748	0.000	-1.080	-0.594
<b>What is your current occupation_Other</b>	-1.1914	0.089	-13.423	0.000	-1.365	-1.017
<b>What is your current occupation_Working Professional</b>	2.3946	0.189	12.642	0.000	2.023	2.766

## Machine Learning/Model building(Contd..)



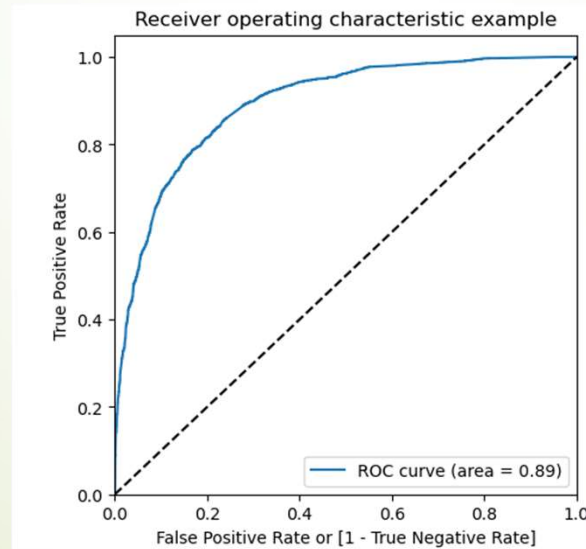
**Optimal cut-off point at 0.35**



# Machine Learning/Model building(Contd..)

	Converted	Converted_Prob	Prospect ID	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted	Lead_Score
0	0	0.059293	302	1	0	0	0	0	0	0	0	0	0	0	6
1	0	0.022673	6087	1	0	0	0	0	0	0	0	0	0	0	2
2	0	0.241976	1033	1	1	1	0	0	0	0	0	0	0	0	24
3	0	0.153273	7656	1	1	0	0	0	0	0	0	0	0	0	15
4	1	0.752513	3241	1	1	1	1	1	1	1	1	0	0	1	75

Assigning lead score to the leads based on optimal cut off of 0.35





# Model Evaluation & Metrics

- Model was evaluated using Precision Recall cut-off of 0.45 which reduced both sensitivity & Recall scores. This was not ideal for the nature of business problem had on hand & hence continued based on 0.35 threshold only.
- Made predictions on test set

## Train data evaluation metrics

Accuracy : 81%

Sensitivity : 81%

Specificity : 80%

Precision: 72%

Recall: 81%

## Test data evaluation metrics

Accuracy : 80%

Sensitivity : 81%

Specificity : 80%

Precision: 72%

Recall: 81%

## Final Model

-0.2785 - 1.6051xDo Not Email + 1.0675xTotal Time Spent on Website - 0.8760xLead  
Origin\_Landing Page Submission + 2.8909xLead Origin\_Lead Add Form + 1.0952xLead  
Source\_Olark Chat + 3.1667xLead Source\_Welingak Website - 1.1395xLast Activity\_Converted  
to Lead - 1.2742xLast Activity\_Olark Chat Conversation + 1.3083xLast Activity\_SMS Sent +  
1.3908xLast Activity\_Unsubscribed - 0.8366xSpecialization\_Other - 1.1914xWhat is your  
current occupation\_Other + 2.3946xWhat is your current occupation\_Working Professional

## Feature Importance

Features importance with respective percentages are as under

Lead Source_Welingak Website	100.000000
Lead Origin_Lead Add Form	91.291311
What is your current occupation_Working Professional	75.617920
Last Activity_Unsubscribed	43.917779
Last Activity_SMS Sent	41.314677
Lead Source_Olark Chat	34.584874
Total Time Spent on Website	33.709718
Specialization_Other	-26.417332
Lead Origin_Landing Page Submission	-27.661534
Last Activity_Converted to Lead	-35.984065
What is your current occupation_Other	-37.624009
Last Activity_Olark Chat Conversation	-40.236284
Do Not Email	-50.687170

# Recommendations

Top 3 features contributing to lead conversions are

1. Lead Source\_Welingak Website
2. Lead Origin\_Lead Add Form
3. What is your current occupation\_Working Professional

Top 3 features to be focussed more for having significant leads conversions are

1. Do Not Email
2. Last Activity\_Olark Chat Conversation
3. What is your current occupation\_Other

## Recommendations(Contd..)

- Lead Origins like 'API' & 'Landing Page Submission' have significant impact on lead conversions.
- Lead Sources like 'Google', 'Direct Traffic', 'Reference', 'Olark Chat' & 'Organic Search' have significant impact on lead conversions.
- Those who have opted for email have significant impact on lead conversions.
- If the Last Activity is either 'SMS sent' or 'Email opened', then they have significant impact on lead conversions.
- Those who have mentioned specialization as 'Others' have significant impact on lead conversions.
- Those who are 'Unemployed' have significant impact on lead conversions.
- Those who have not opted for 'A free copy of Mastering The Interview' have significant impact on lead conversions
- Those who have visited the website atleast 5 times have significant impact on lead conversion rate.
- Those who have spent total time on website of around 1200 units have significant impact on lead conversion rate.
- Those who have viewed Average number of pages on the website of atleast 3 have significant impact on lead conversion rate

## Recommendations(Contd..)

Few areas they can look into to create more potential leads are as under

- a. *E-mail reminder campaigns*
- b. *Improvise Marketing strategies around preferred lead sources like 'Google', 'Direct Traffic', 'Reference', 'Olark Chat' & 'Organic Search'.*
- c. *Referral programs with eye-catching incentives.*
- d. *Target customers whose occupation mentioned as 'others'(to understand more about them & filter out who among these might be really interested in pursuing the learning/course)*



THANK YOU