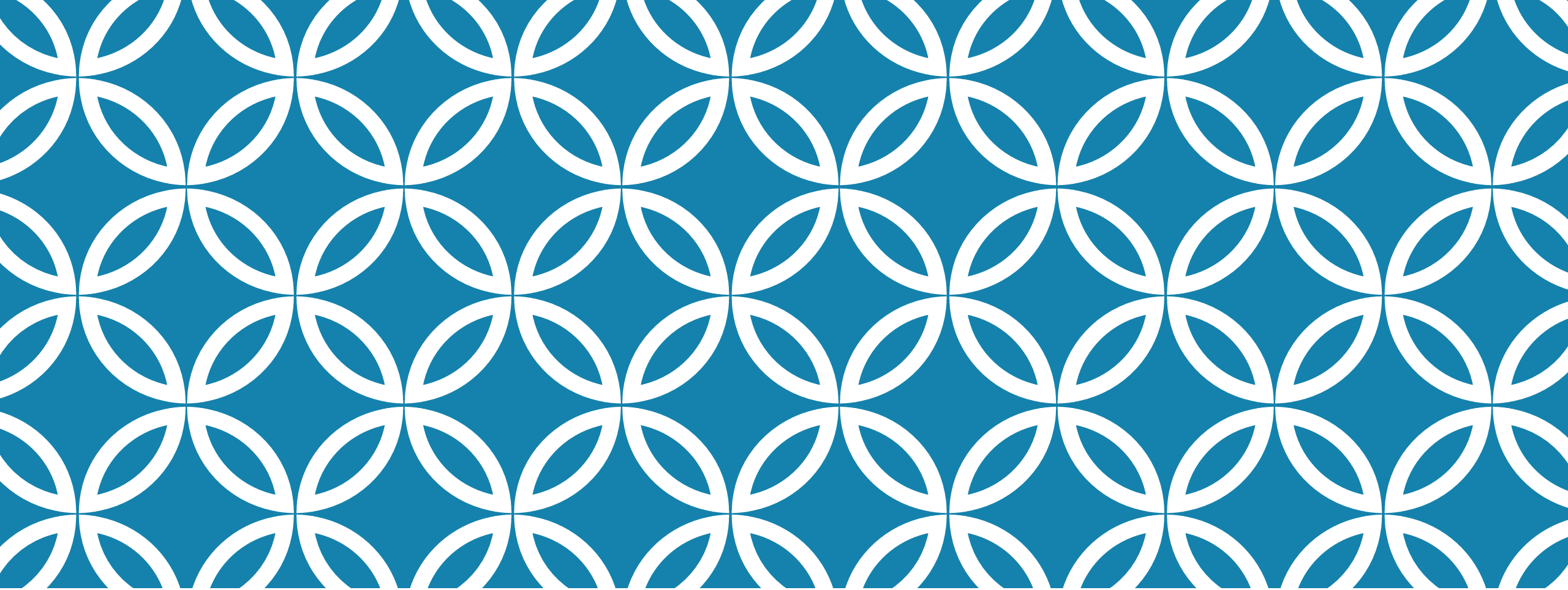


LEAD SCORING CASE STUDY

Kunal Malhan
ACP DS, IIITB

UNDERSTANDING - PROBLEM STATEMENT & OBJECTIVES

X Education is an education company that sells online courses.

Company receive the leads from many sources like their own website, references, through marketing on various websites, search engines like Google, social media sites like YouTube, facebook etc.

Currently, conversion rate for these leads to actual customers is approximately 38%.

Business Objectives for Use-case

- The aim of the study is to identify the leads that are most likely to convert into paying customers.
- Assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- Handling special requirement for adjusting model to support:
  - Identify almost all of the potential leads during 2 months when additional manpower like Interns are available.
  - Identify strategy to minimize phone calls during times when company reaches its target for a quarter before the deadline.

OVERALL APPROACH OF THE ANALYSIS

Understanding Problem Statement and Datasets

Exploratory Data Analysis

- Missing Value Handling
- Data Cleaning and Manipulation

Univariant Analysis (using data visualization)

- Categorical Flags (Yes/No) Variables (using Pie plot / Bar plot etc)
- Categorical Variables (using Bar graphs and Stack graphs etc)
- Numerical Variables (using Distribution plot, Box plots etc.)

Bivariant and Multivariant Analysis

- Categorical vs Categorical Variables (using Bar graphs and Stack graphs)
- Numerical vs Numerical Variables (using correlation heatmap, and pair plot)

Data Preparation

Building Model

Results Evaluation

Assignment of 'Lead Score' to each of lead

MISSING VALUES – IDENTIFICATION & HANDLING

Column Name	%	Handling
How did you hear about X Education	78.46	Dropped
Lead Profile	74.19	Dropped
Lead Quality	51.59	Dropped
Asymmetrique Profile Score	45.65	Dropped
Asymmetrique Activity Score	45.65	Dropped
Asymmetrique Profile Index	45.65	Dropped
Asymmetrique Activity Index	45.65	Dropped
City	39.71	Dropped
Specialization	36.58	New category 'Unknown' for missing values.
Tags	36.29	Restructuring Tags Category, and merged missing values with Low Chances Category.
What matters most to you in choosing a course	29.32	Dropped
What is your current occupation	29.11	Proportionately distribute missing values.
Country	26.63	Dropped
TotalVisits	1.48	Mean for missing values.
Page Views Per Visit	1.48	Mean for missing values.
Last Activity	1.11	Dropped, due to similar column already exist.
Lead Source	0.39	Mode for missing values.

High Missing Values	
Mid Missing Values	
Low Missing Values	

Missing Values includes the records that have 'Select' as value.

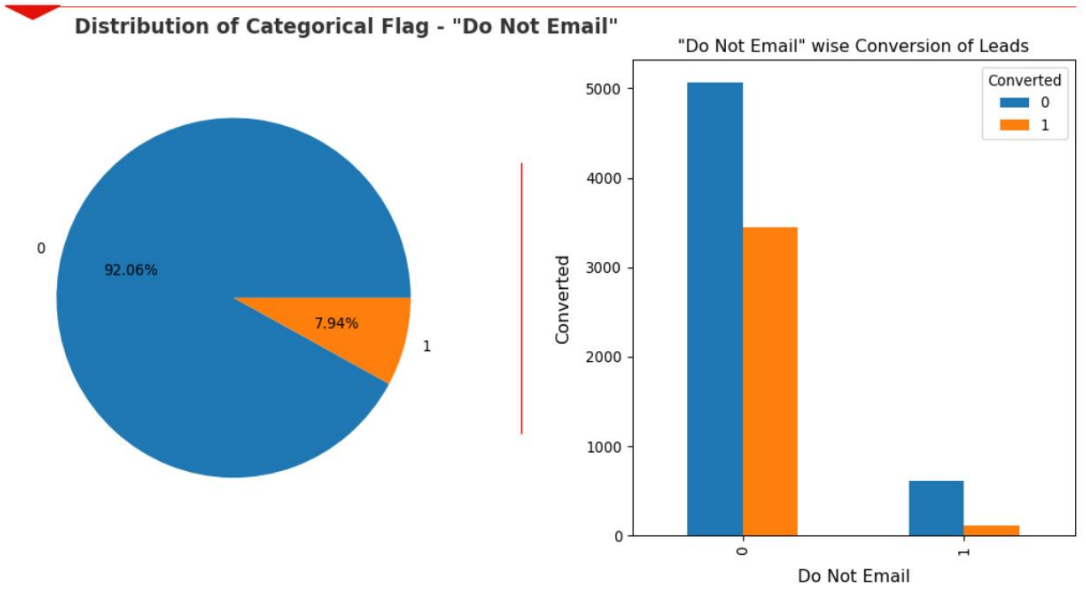
DATA CLEANING ACTIVITY

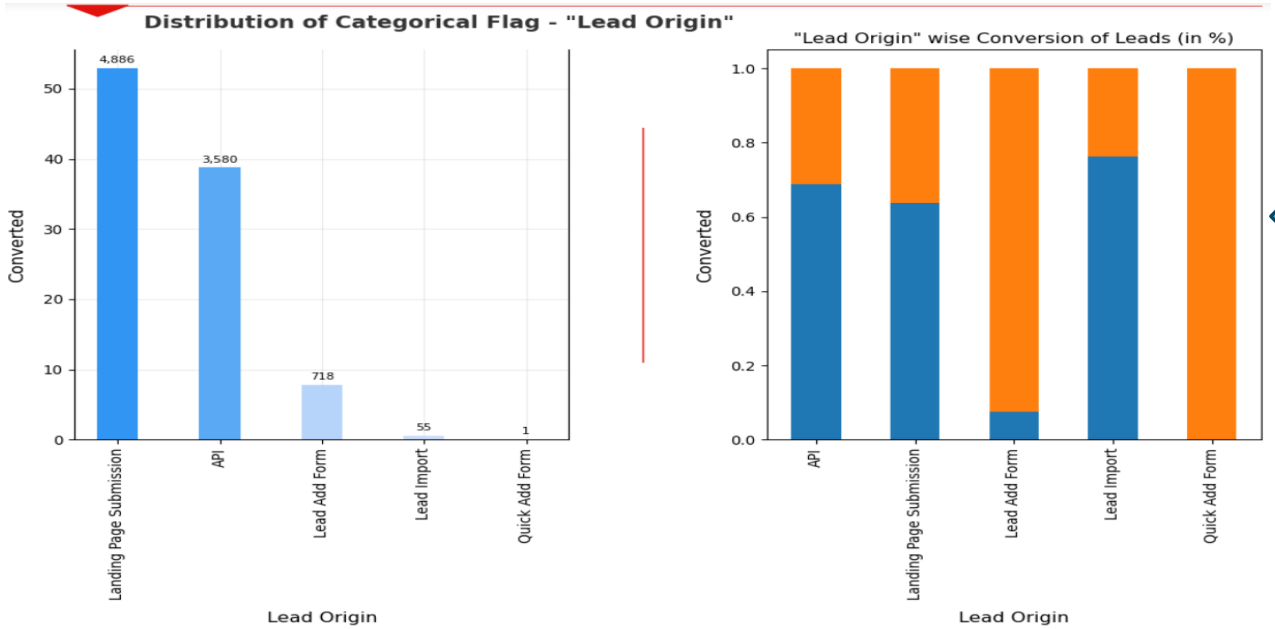
Variable	Action
'Specialization'	New categories created: • ‘Management’: All values that contains ‘Management’ are clubbed. • ‘Business’: All values that contains ‘Business’ are clubbed. • ‘Others’: All values that are not related to ‘Management’ and ‘Business’.
‘Tags’	New variable ‘Tags_Conversion_Chances’ derived: • ‘LOW_CONVERSION_CHANCES’: Clubbed categories options with conversion upto 25%. • 'MODERATE_CONVERSION_CHANCES’: Clubbed categories options with conversion 25-80% • 'HIGH_CONVERSION_CHANCES’: Clubbed categories options with conversion above 80%.
'occupation'	• Clubbing low frequency values of 'Housewife’, 'Businessman’ to make ‘Others’
‘Lead Source’	• ‘google’ is clubbed with ‘Google’. • All low frequency options except Top 5 ('Google', 'Direct Traffic', 'Olark Chat', 'Organic Search', 'Reference'), merged as 'Others’.
'Last Notable Activity'	• All low frequency options except Top 3 ('Modified', 'Email Opened', 'SMS Sent'), merged as 'Others’.
'TotalVisits'	• Dropped due to high correlation (>70%) with ‘Page Views Per Visit’.
'Page Views Per Visit'	• Outliers in ‘Page Views Per Visit’ are replaced with 99 percentile.

UNIVARIANT ANALYSIS

Categorical Flags (Yes/No) Variables

Dropping columns where single value has > 99% occurrences.





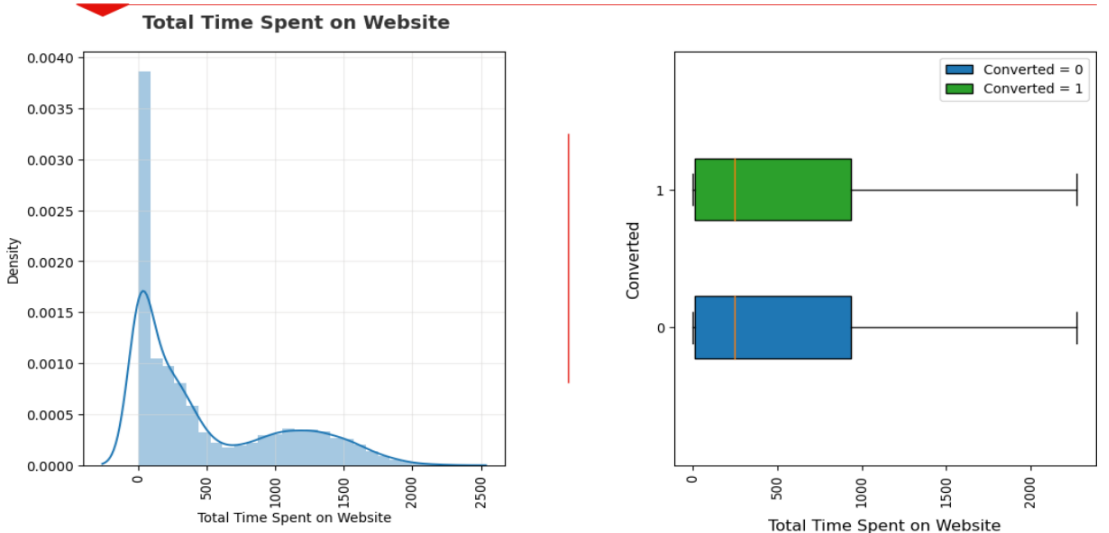
The P-Value of ChiSq Test: 1.3384599721779416e-38
Independent variable - "Do Not Email" is correlated to target variable - "Converted"

Categorical Variables

Clubbing low frequency categories in 'Lead Source', 'Last Notable Activity'.

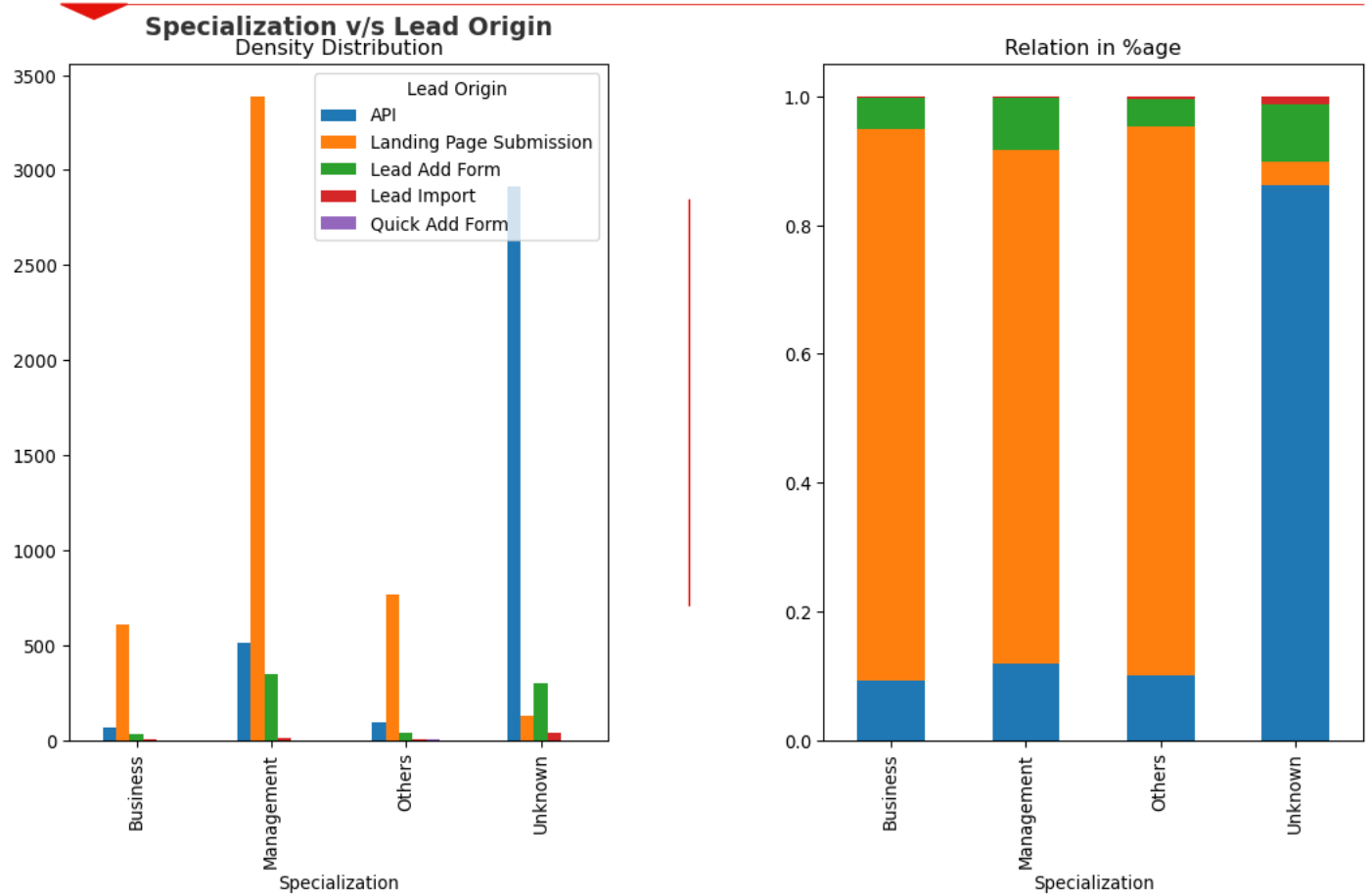
Numerical Variables

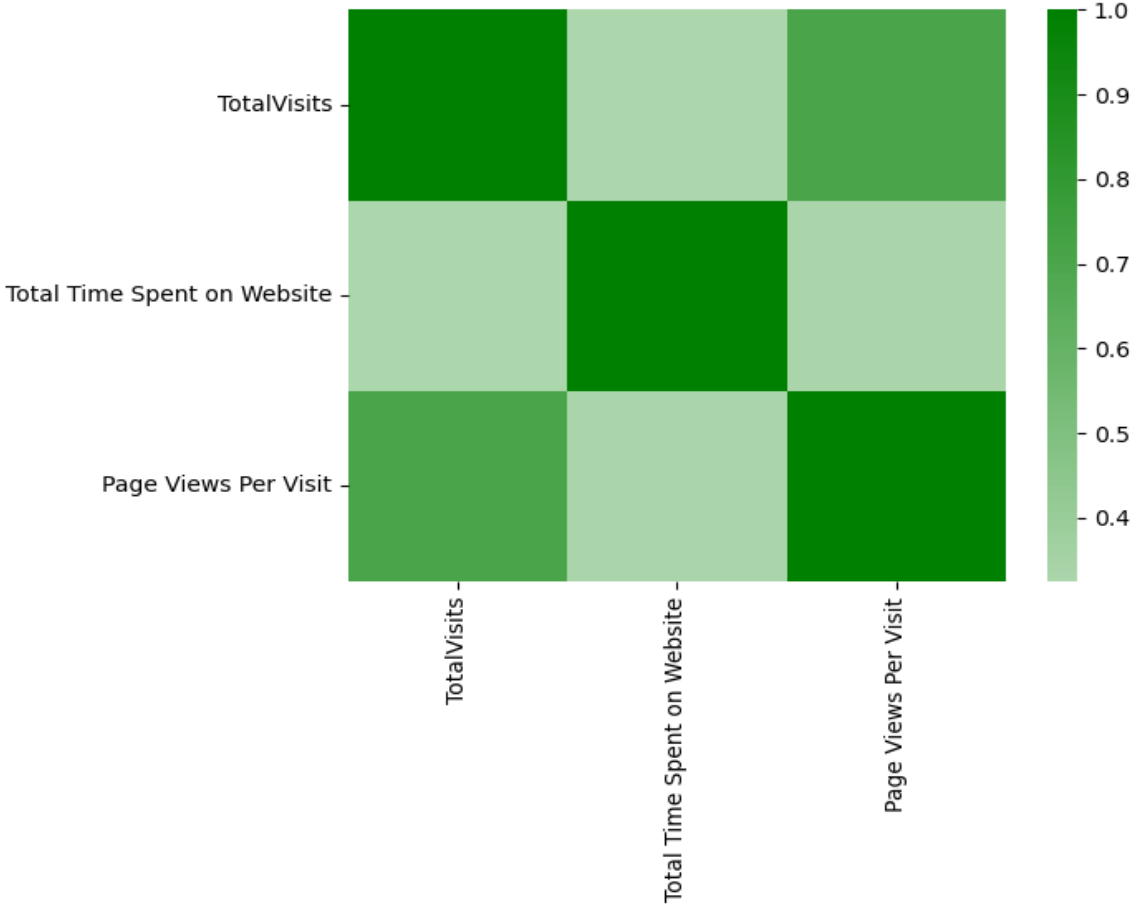
Outliers handling for columns 'TotalVisits', ‘Page Views Per Visit’.



MULTIVARIANT ANALYSIS

- Dropping column 'TotalVisits' due to its high correlation with ‘Page Views Per Visit’.
- 'Direct Traffic’, 'Olark Chat’, and ‘Reference’ in 'Lead Source' mostly correspond to 'Landing Page Submission’, 'API’, and ‘Lead Add Form’ respectively in 'Lead Origin’.
- Most of 'Unemployed' members from lead have 'Management' as 'Specialization'.





DATA PREPARATION

Replacing categorical variables with dummy variables.

Instead of using first_drop, selective column made from dummy variables has dropped.

Making column 'Lead Number' as index and dropping as column.

Splitting the Data into Training and Testing Sets in 70-30 ratio.

Rescaling the Numerical Features using Standard Scalar.

Finding correlation/Heatmap between all variables, and dropping variables with high correlation > 70%.

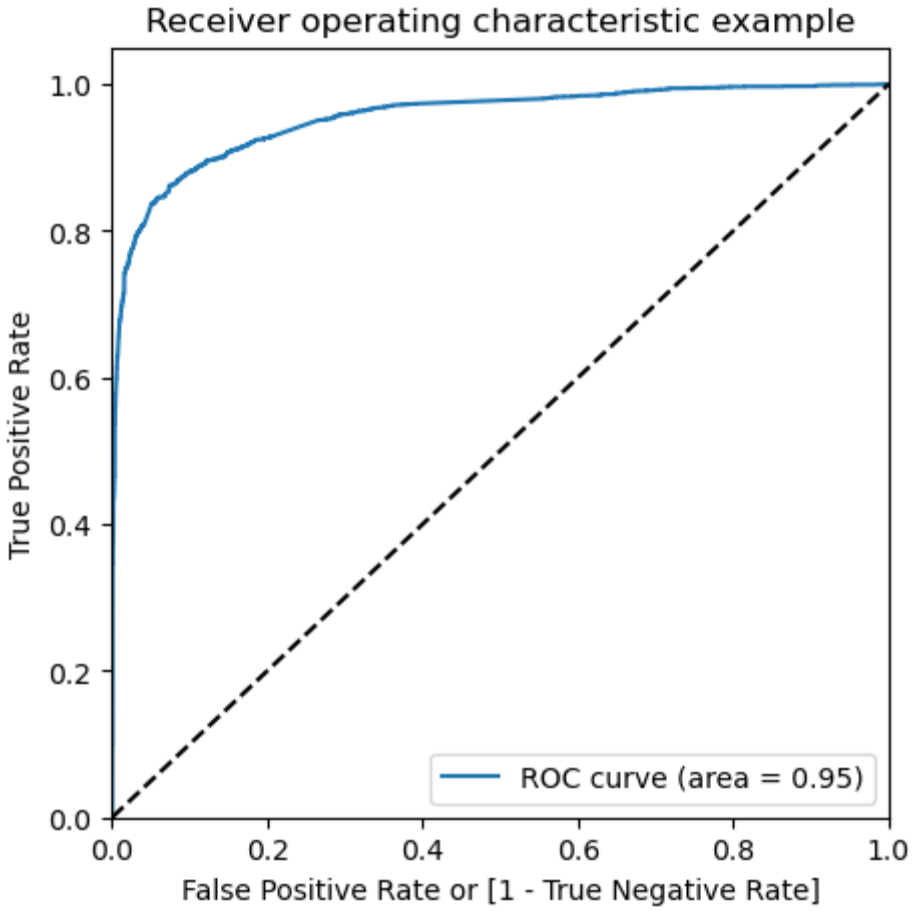
BUILDING MODEL

Final dataset after cleaning and preparation consists of 19 variables.

In first logistic regression model using statsmodels, it is found that there exist variables with high p-value.

With multiple experiments using automatic feature selection (RFE), adequate 11 features have been selected, where all variables have p-value < 0.05 and VIF in range.

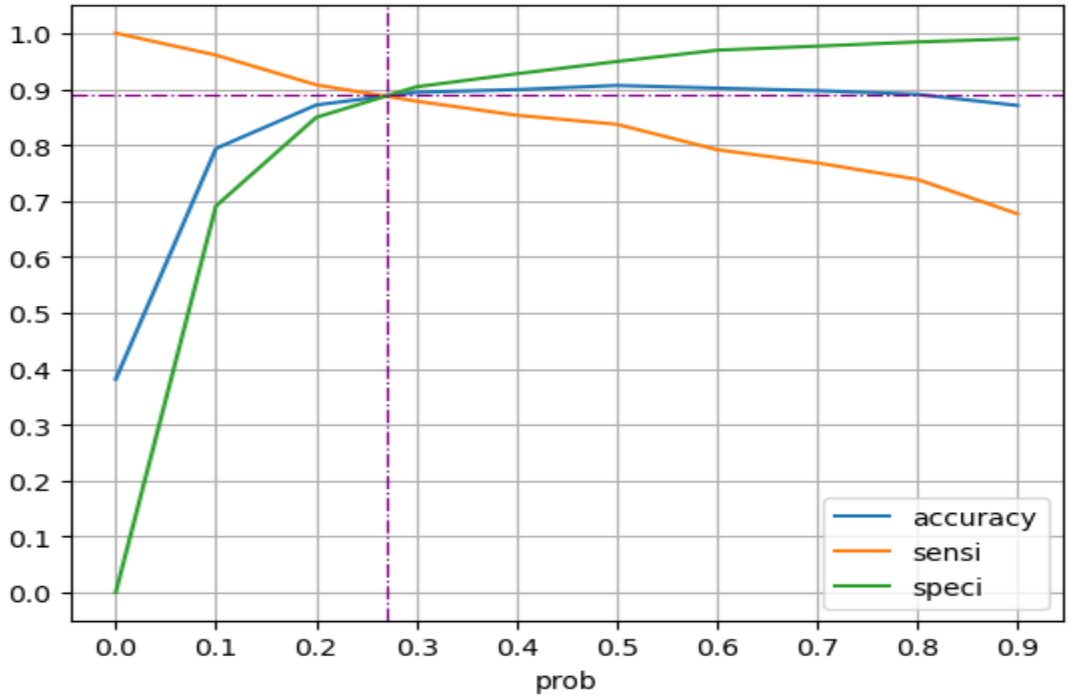
ROC plot show ROC curve area is 0.95, which is pretty good.



ACCURACY, SENSITIVITY, AND SPECIFICITY BALANCE V/S PRECISION, AND RECALL BALANCE

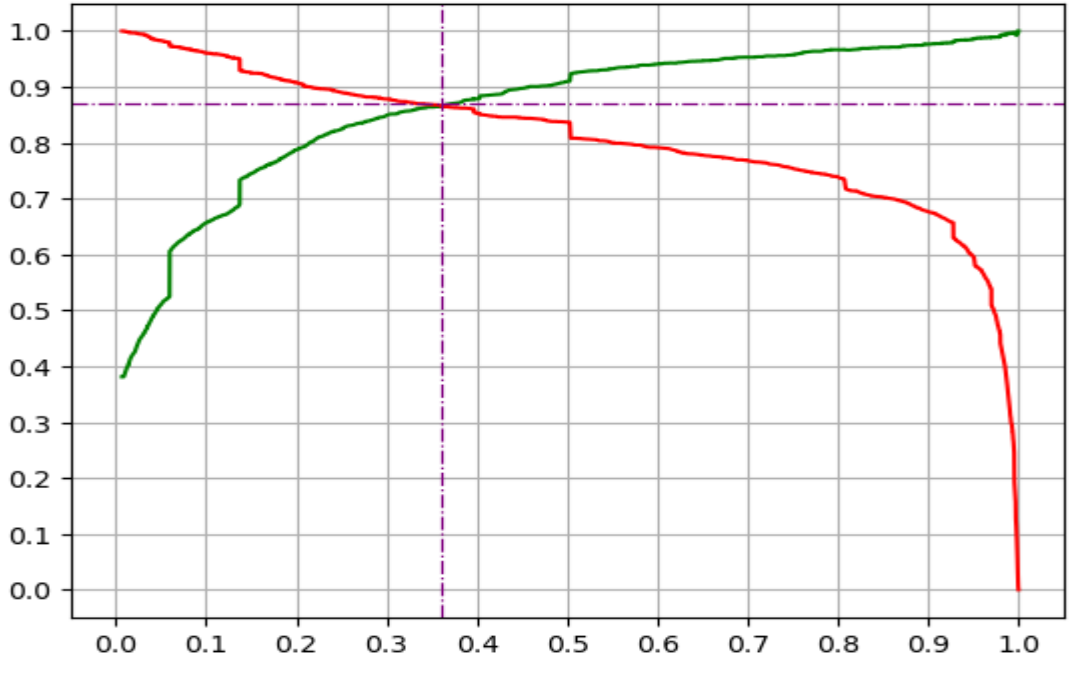
Accuracy, Sensitivity, and Specificity Balance

Cut-Off – 0.27



Precision, and Recall Balance

Cut-Off – 0.36



Cut-off of 0.36 is chosen to convert the conversion probabilities to actual Converted Flag.

MODEL RESULTS

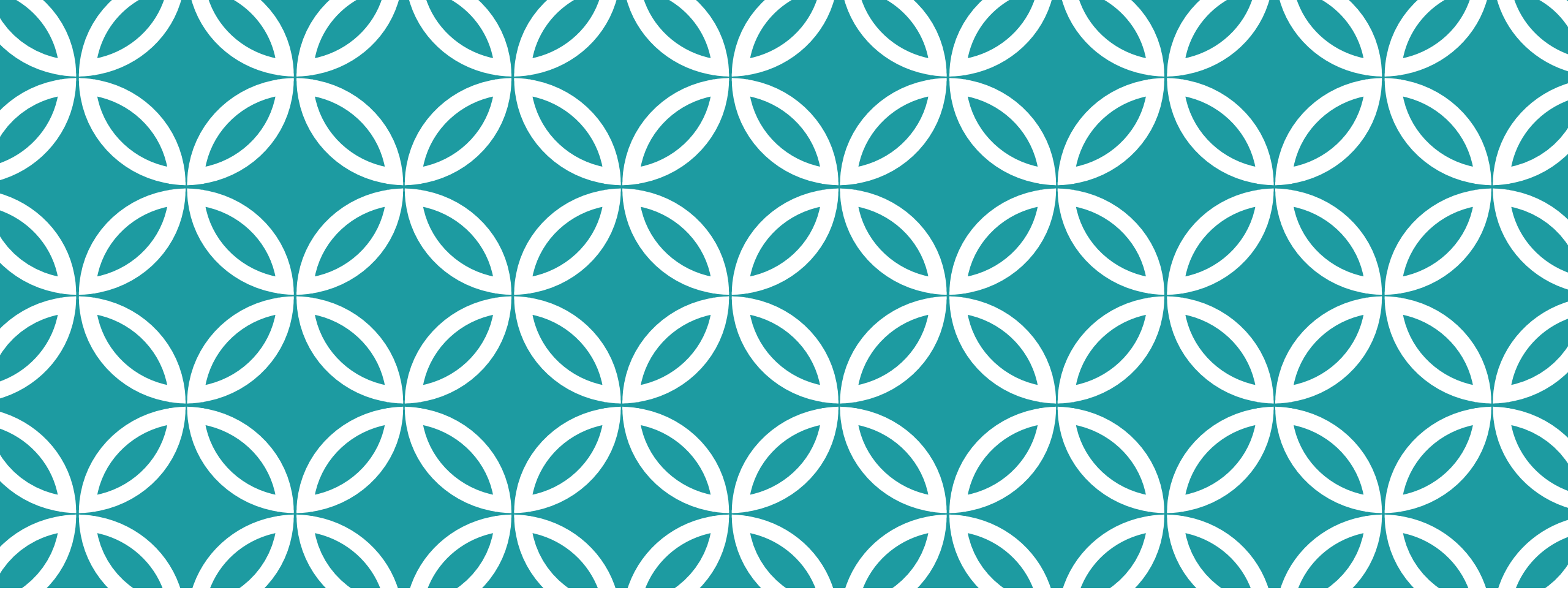
	Accuracy, sensitivity, and specificity balance [Train Data]	Precision, and Recall Balance [Train Data] (Finally selected)	On Test Data
Optimal Cut-off	0.27	0.36	
Confusion Matrix • True Positives (TP) • True Negatives (TN) • False Positives (FP) • False Negatives (FN)	 • 2181 • 3571 • 431 • 285	 • 2134 • 3669 • 333 • 332	 • 949 • 1547 • 130 • 146
Accuracy	0.8893	0.8972	0.9004
Sensitivity	0.8844	0.8654	0.8667
Specificity	0.8923	0.9168	0.9225
Precision	0.8350	0.8650	0.8795
Recall	0.8844	0.8654	0.8667

OUTCOMES FOR BUSINESS (MOST EFFECTIVE FEATURES)

- Historical responses by leads like 'Will revert after reading the email', 'Interested in Next batch', and indications by sales team based on interaction with leads like 'Lost to EINS', 'Closed by Horizzon', 'Lateral student', provide very good conversion rate at later stage.
- The audience/leads that are approached through SMS have provided positive outcomes, so it increasing reach to audience/leads through SMS should be increased.
- It is observed that leads that spend good time on X Education's website have higher chances of converting to customer. So, X Education can work on enhancing feature and information on their website, and make it more interactive.
- Leads that open X Education Emails are also good candidates to target, so title for these Emails can be made more innovative, make user to open to know more about programs.

RECOMMENDATIONS FOR SPECIAL PERIODS

Aggressive Outcomes (10 additional Interns/Resources)	Team busy with other work (Only essential)
Model Adjustment and Enhancement • Decrease the model cut-off probability to have **higher 'Recall'**, which will cover most of leads that have any chance to be converted. • Rebuilding model to use latest statistic and make new if need arises.	**Model Adjustment and Enhancement** • Increase the model cut-off probability to have **higher 'Precision'**, which will return more relevant results, rather than irrelevant results. • Rebuilding model to use latest statistic and make new if need arises.
Additional Strategies • Call to leads with higher lead score and move toward leads with lower lead score. • Increase reachability to leads through SMSs/Emails. • Enhancing feature and information on their website, and make it more interactive.	**Additional Strategies** • Follow-up through automated SMS service to focused audience. • Enrich Website to support chatbot and multi-level questioning to get more in-depth information of landed audience and identify only highly interested audience. • Whenever phone service is used than choose the lead with higher 'Lead Score' at any stage.



THANK YOU