



Lead Scoring Case Study



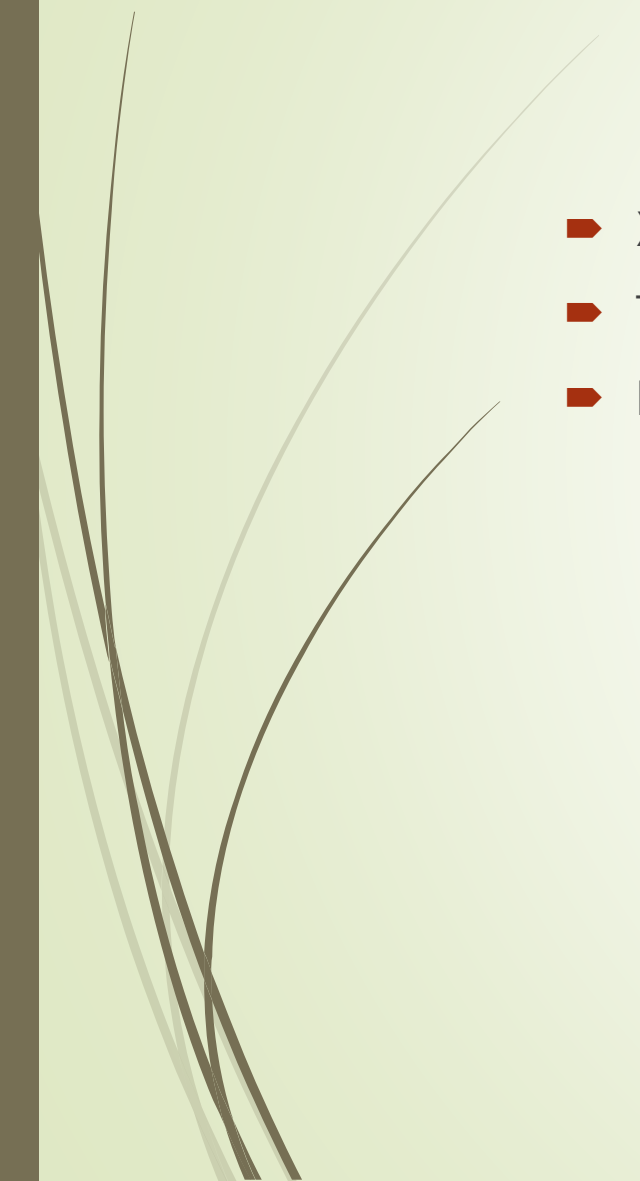
Problem Statement



- Online courses are offered by X Education to business professionals.
- X Education receives a large number of leads, but its lead conversion rate is very low. For instance, if they receive 1000 leads in a day, only approximately 300 of them will actually convert.
- The organization wants to find the "Hot Leads," or leads with the highest potential, in order to increase the efficiency of this procedure.
- If they are able to locate this group of prospects, the lead conversion rate should increase because the sales staff will be concentrating more on contacting potential leads rather than calling everyone.



Business Goal

- X education wants to learn about the most promising leads.
 - The aim is to create a model that detects hot leads for that purpose.
 - Deploying and using the model for best results during future activities.
- 



Approach Technique

➤ Data Comprehension

- Understand the data using various functions

➤ Cleaning up Data

- Identifying and removing duplicates
- Removing NaN and missing values
- Column Dropping and Imputation
- Deal with outlier data

➤ Preparing Data for Model Building

- EDA
- Convert Binary Variables
- Create Dummy Features
- Drop Repeated Variables



Approach Technique

➤ **Model Building**

- Create Test-Train Split
- Scale Features

➤ **Model Evaluation**

- Feature Selection
- Model Assignment using StatsModel
- VIF value check
- Model Performance check and ROC Curve Plot Creation
- Model Tuning with Cutoff point
- Accuracy, Sensitivity, Specificity, false positive rate, Positive predictive value, and Negative predictive value on train set checks
- Precision and Recall Trade-off

➤ **Making Predictions on Test Set**

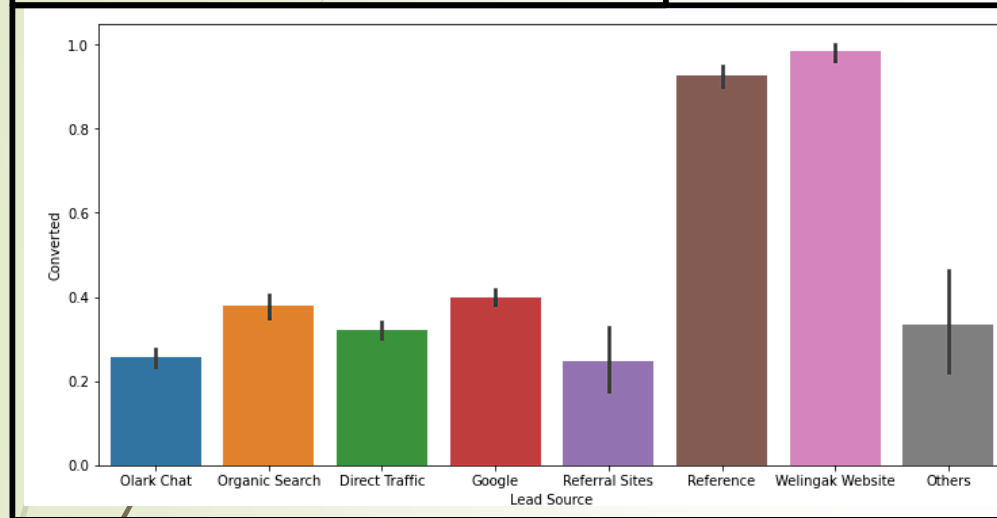


Data Comprehension and Clean up

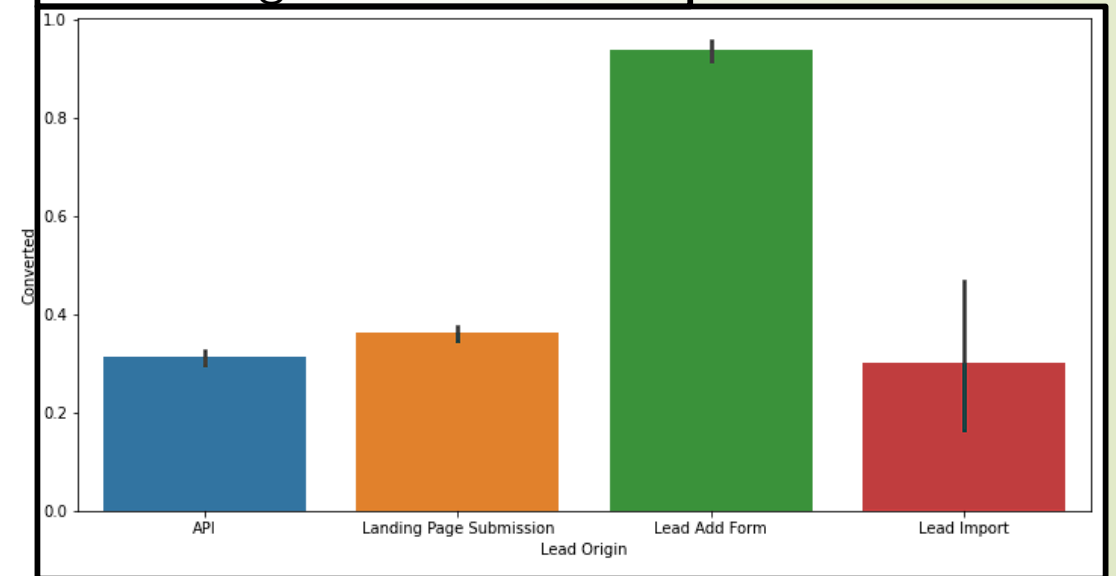
- Data Shape Details = 37 Rows, 9240 Columns
- Converting Select values from Data to Null
- Huge number of increases in null values can be seen in various columns now. Especially in 'Lead Profile' and 'City'
- Dropping columns with high Null values
- Impute values in columns where needed as per analysis
- Verifying Unique values on all columns with unique IDs
- Dropping Prospect ID and Lead Number which do not have any importance for performing our analysis
- Dropping Repeated Variables
- Working with Outliers

Exploratory Data Analysis

Lead Source vs Converted

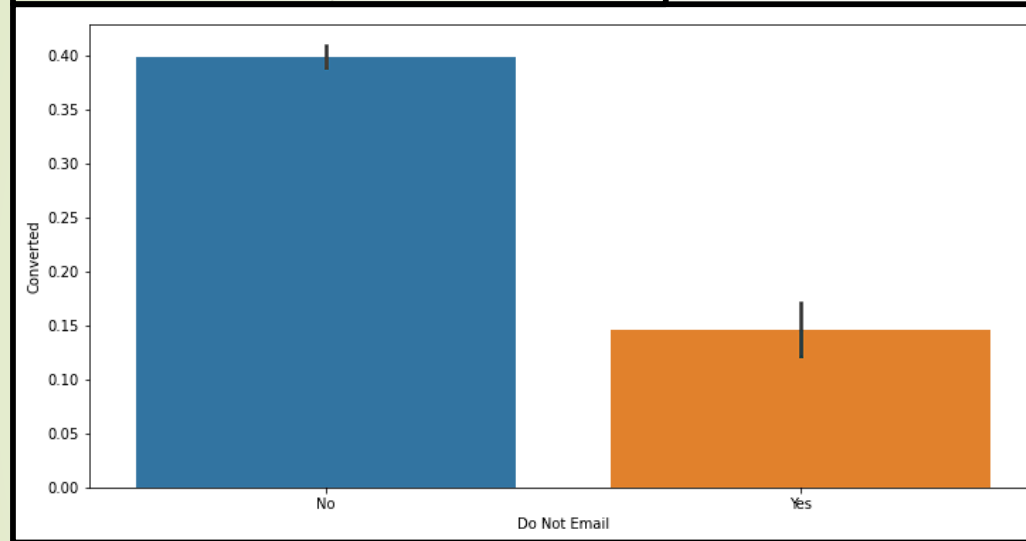


Lead Origin vs Converted

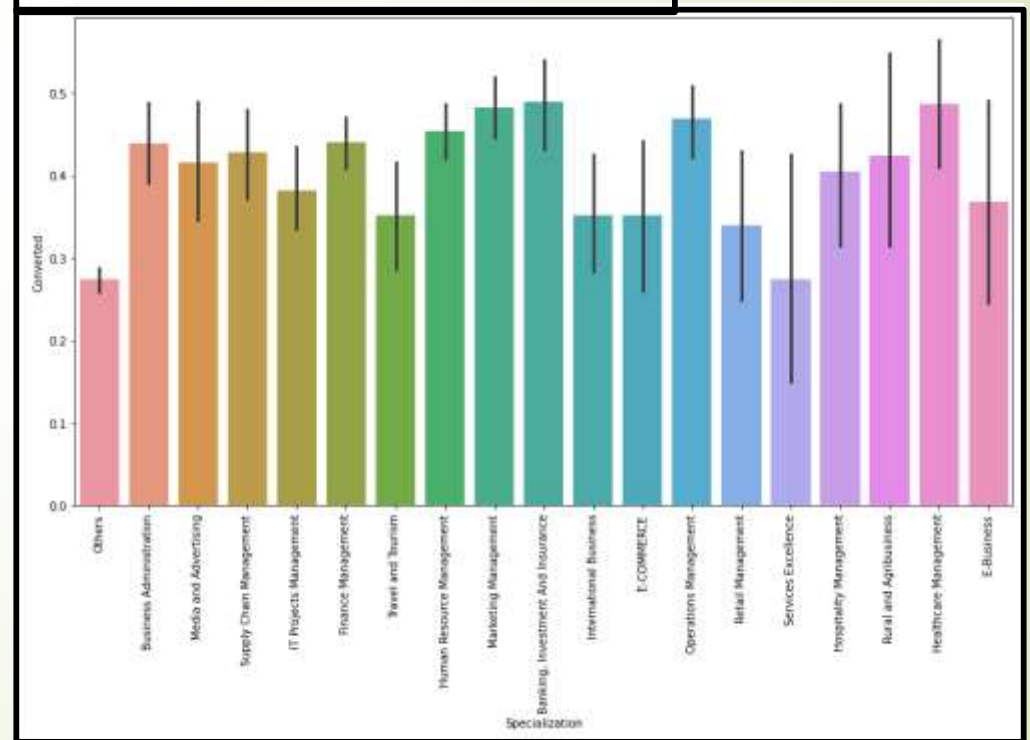


Exploratory Data Analysis

Do Not Email vs Converted

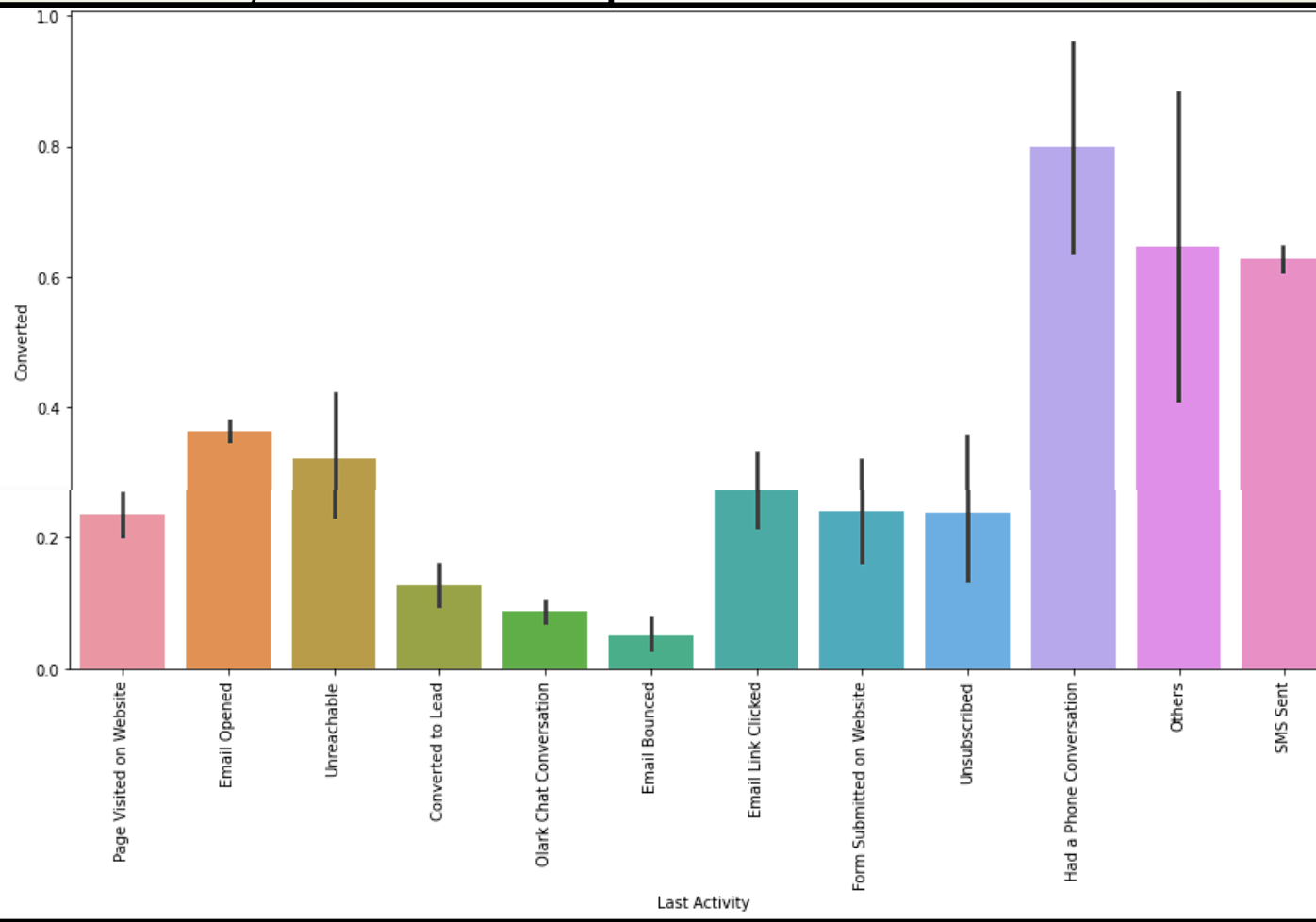


Specialization vs Converted



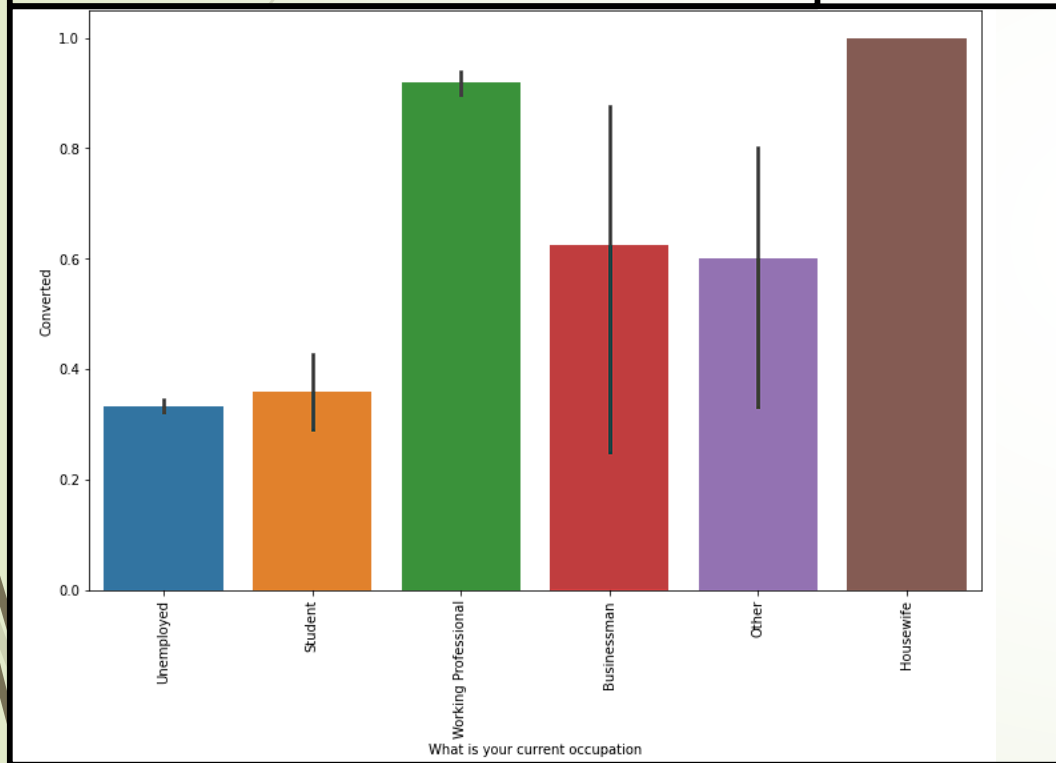
Exploratory Data Analysis

Last Activity vs Converted

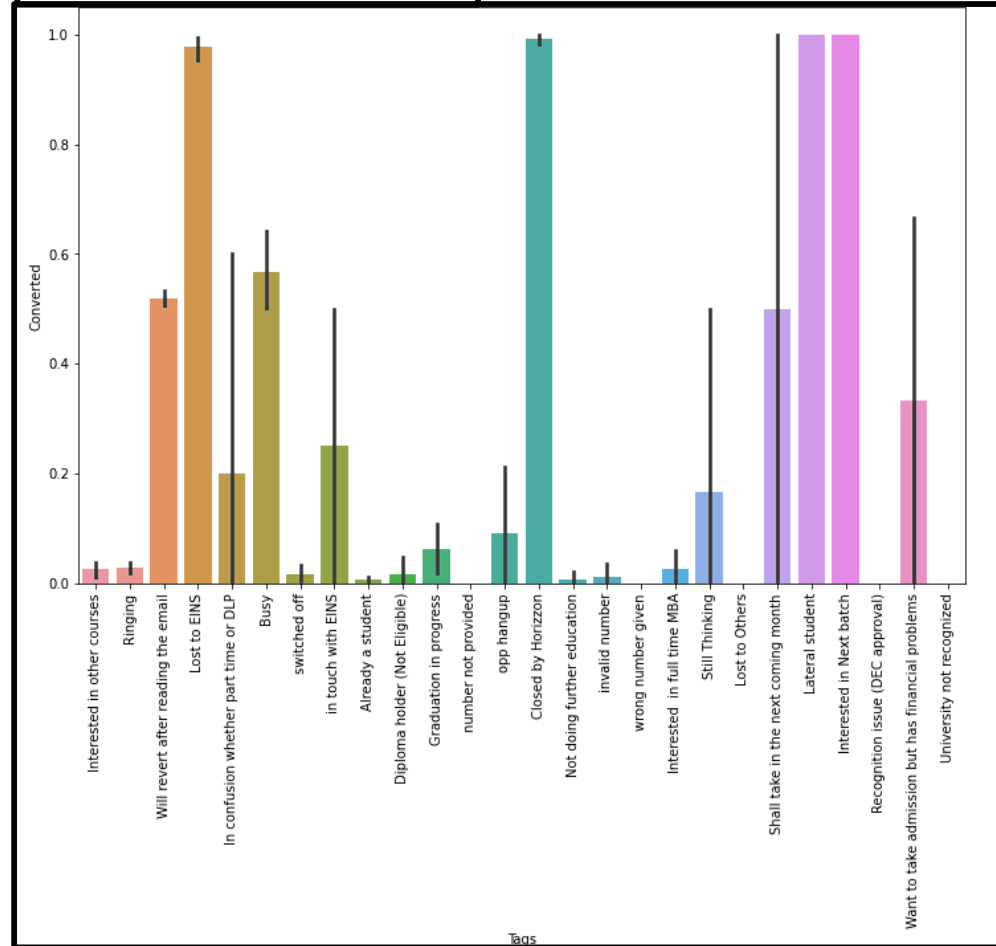


Exploratory Data Analysis

What is Your Current Occupation vs Converted

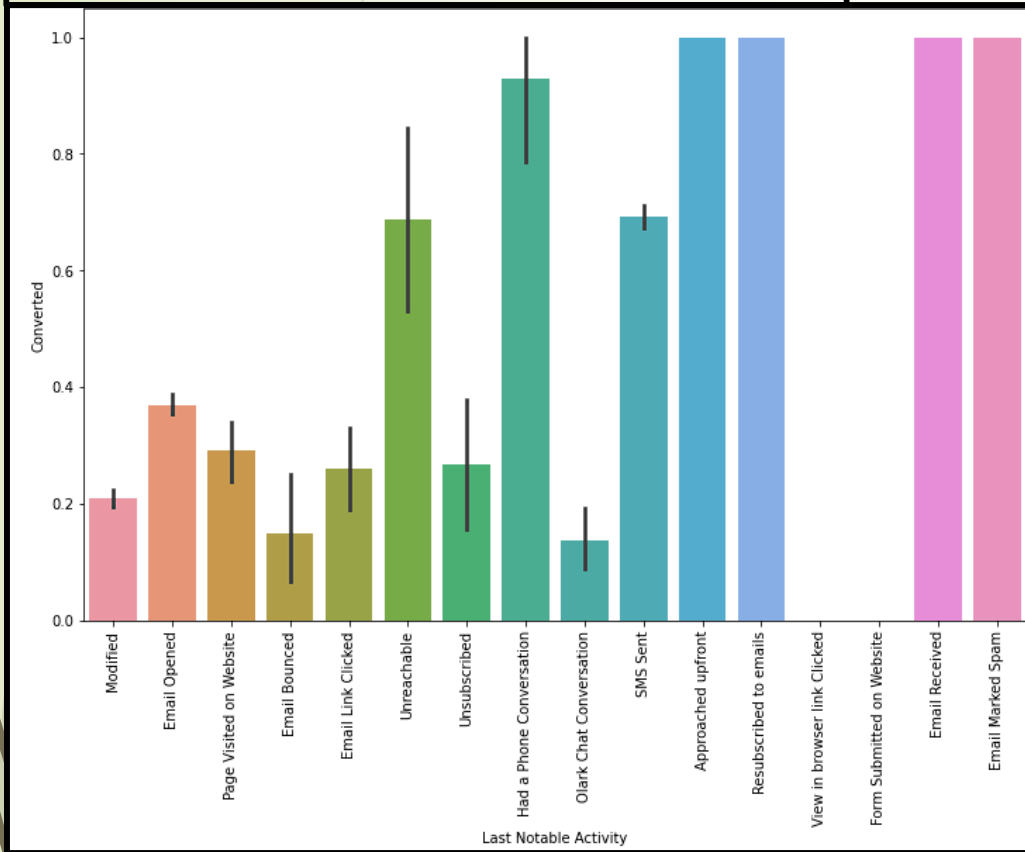


Tags vs Converted

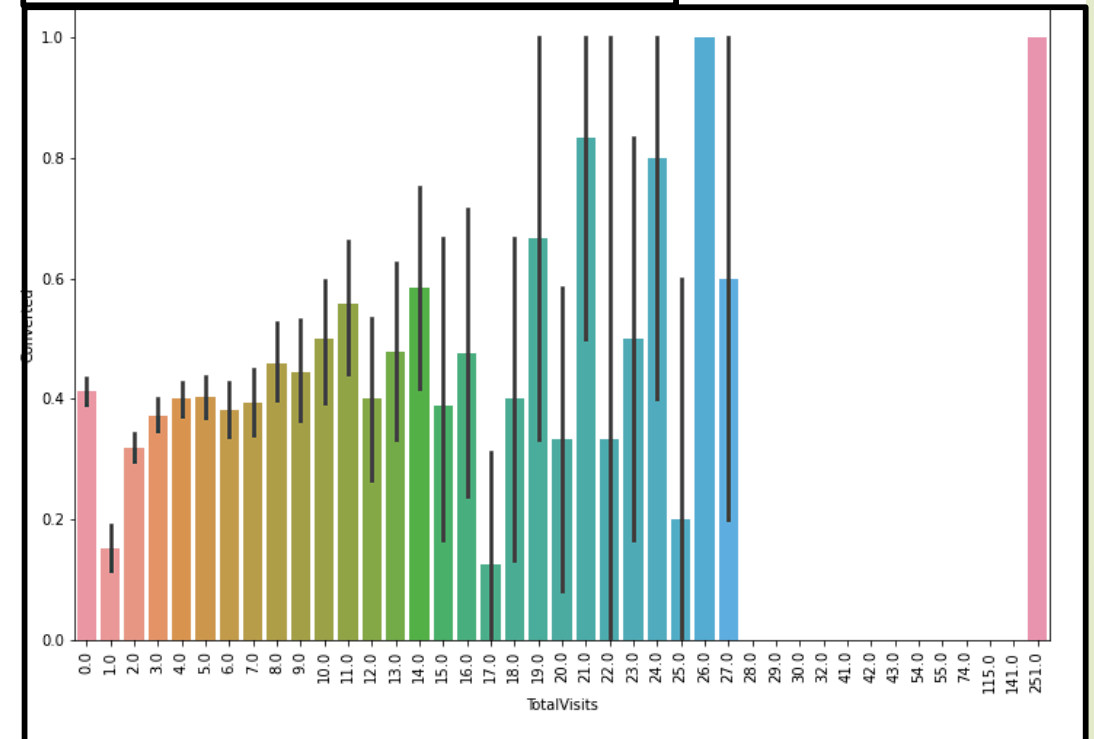


Exploratory Data Analysis

Last Notable Activity vs Converted

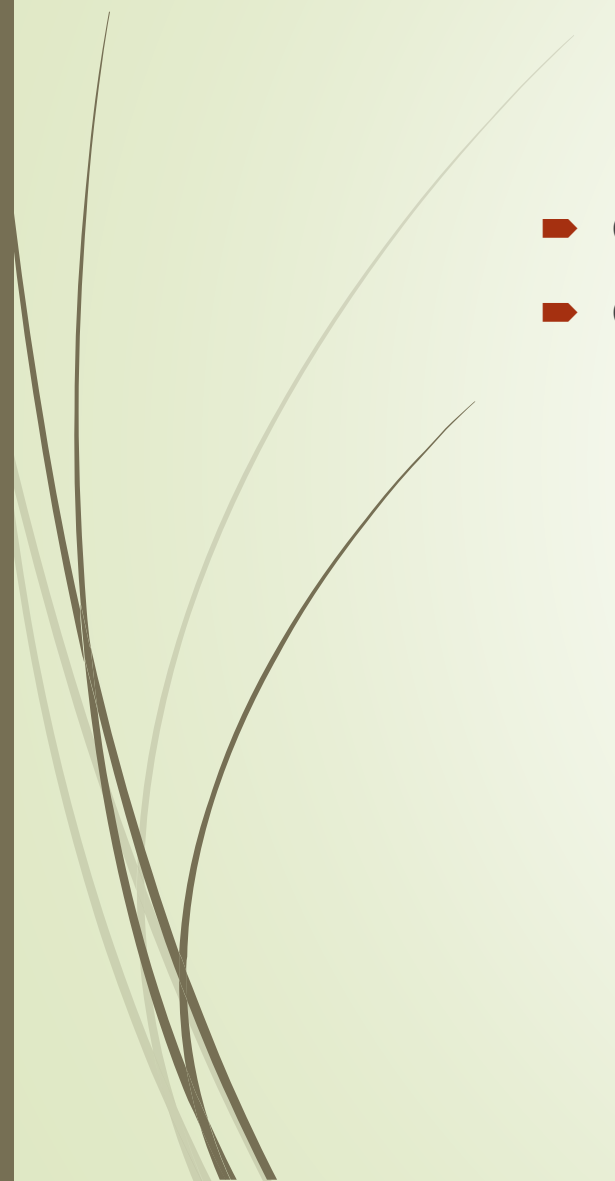


Total Visits vs Converted





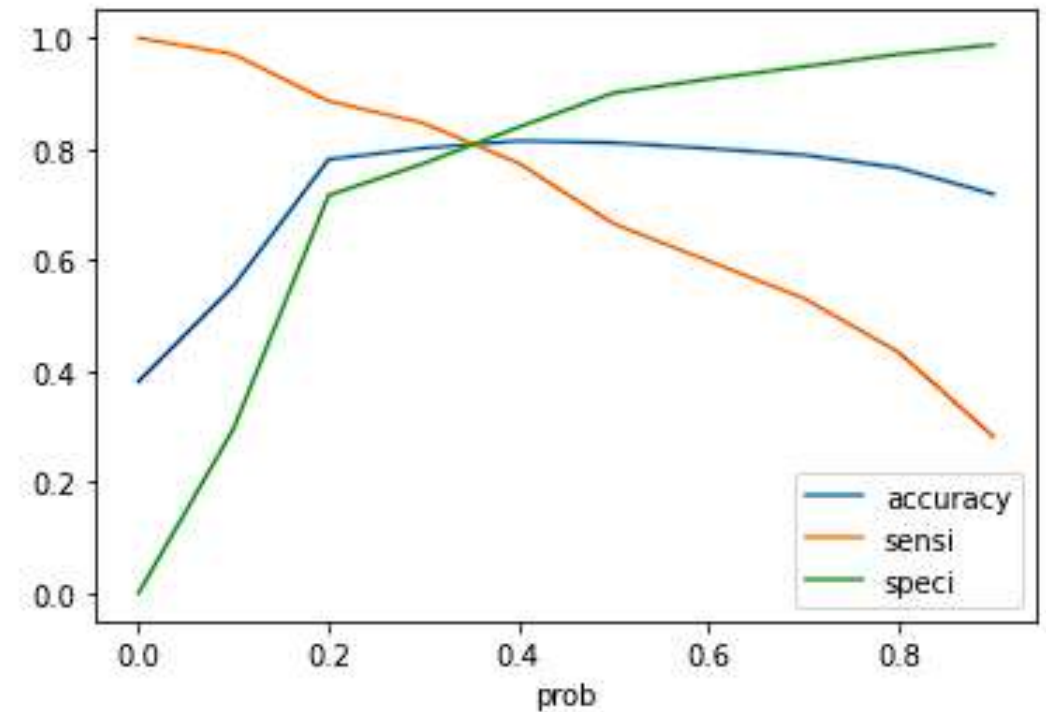
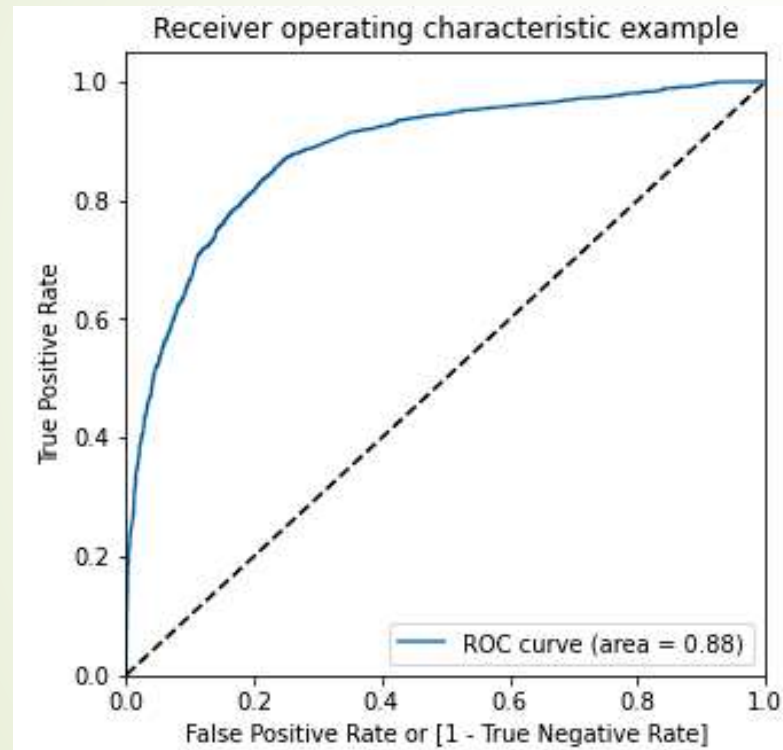
Transformation of Data

- Converting Binary Variables
 - Creating Dummy Variables
- 

Model Building

- ▶ Creating a Training and Test Data-Split
- ▶ Properly Scaling the Features
- ▶ The first split was created using Train Size = 0.7 and Test Size = 0.3
- ▶ Feature Scaling using RFE
- ▶ Assessing the Model using StatsModel
- ▶ Checking VIF values. Output for all features having less than 5
- ▶ Creating a dataframe with the actual Converted flag and the predicted probabilities
- ▶ Confusion Matrix = $\begin{bmatrix} 3605 & 397 \\ 825 & 1641 \end{bmatrix}$
- ▶ Overall Accuracy = 81%
- ▶ Sensitivity = 66%, Specificity = 90%, False Positive = 10%, Positive = 80%, Negative = 81%

ROC Curve



- The second graph demonstrates that the ideal cutoff is around at 0.3

Making Predictions on the Test Set

Final Test Predictions (head)

	Converted	Lead Number	Converted_Prob	final_predicted
0	1	4269	0.749414	1
1	1	2376	0.915483	1
2	1	7766	0.957115	1
3	0	9199	0.108475	0
4	1	4359	0.897277	1

- Overall Accuracy = 80%, , Sensitivity = 83%, Specificity = 77%
- Confusion Matrix = $\begin{bmatrix} 1303, & 374 \\ 179, & 916 \end{bmatrix}$



Model Summary

Training Data		
Accuracy	Sensitivity	Specificity
0.80	0.84	0.77

Test Data		
Accuracy	Sensitivity	Specificity
0.80	0.83	0.77



Final Interpretation

- As per the research and study, the following are the factors that mostly affected the prospective purchasers – Total Number of Visits & Total time spent on the website
- For the above-mentioned scenarios, Lead source was Olark Chat Welingkar Website, Google or Organic Search
- When the last activity occurred via SMS or Olark Chat
- If the Lead add format is the lead origin used.
- When the current Occupation in sequence is either:
 1. Working Professionals
 2. Student
 3. Unemployed
 4. Other
- With this available data, X Education can increase the number of prospective customers and also help them think more positively about accepting the X Education courses