Lead Scoring Case Study

**Presented by:**

Nilesh Thawase

Yuthika Bhandari

Agenda

- Goal of Case Study

- Steps Performed to create ML Model

- Visualization and Observation

- Conclusion

Goal of Case Study

- 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 X Education to target potential leads. A higher score would mean that the lead is hot.

- Improve target lead conversion rate from 30% to around 80%.

- There are some more problems presented by the company which model should be able to adjust to if the company's requirement changes in the future.
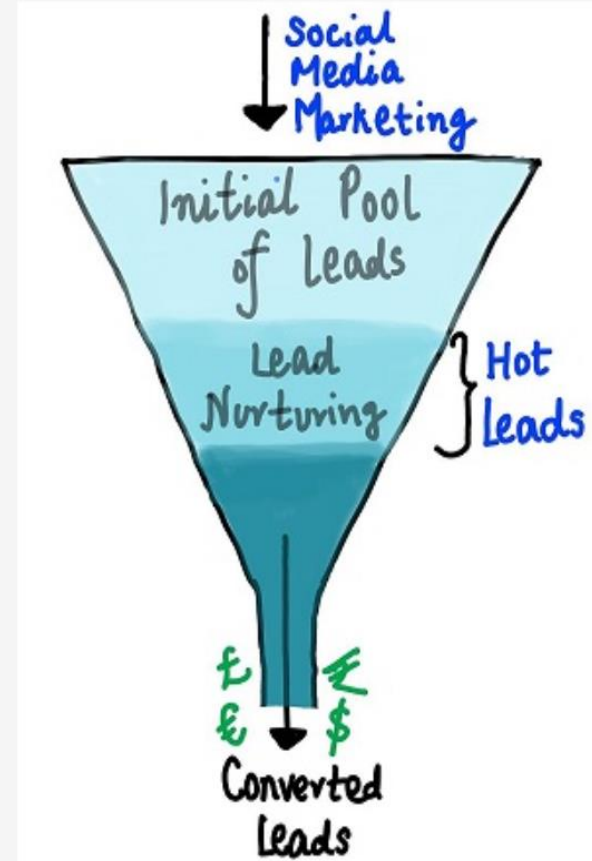
Lead Conversion Process - Demonstrated as a funnel

Steps Performed to create ML Model

- Importing the data and understanding
- Data Cleaning
  - Missing Value Treatment
  - Outlier Treatment
- Numerical Column Multicollinearity check
- Categorical Variable Imputation
- Dropping unnecessary columns
- Creating Dummy Variables for Categorical Columns

- Model Building
  - Train and Test Split
  - Feature Scaling
  - Feature Selection from RFE
  - Creating a Model
  - Assessing model with statsmodel
  - Checking VIF
  - Performing Iterations
  - Predicting on Training dataset
  - Plotting ROC Curve
  - Find Optimal Probability cut-off
  - Finding Accuracy, Sensitivity and Specificity
  - Making Predictions on Test dataset
  - Evaluating the metrics for test dataset
- Adding a Lead Score to data frame
- Finding the final Equation

Data Cleaning

- Dataset contains 9240 rows and 37 columns.

- There are a lot of missing values present

- Unique identification columns like 'ProspectID' and 'Lead Number' were dropped because of their insignificance in model building.

- Columns having single value like 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque' have been dropped

- Dropping columns like 'City', 'Asymmetrique Activity Score' and 'Asymmetrique Profile Score' as they seem insignificant.

- Many columns in the dataset have 'Select' value, we replace it by 'NaN'

Data Cleaning

- Removing columns like 'How did you hear about X Education', 'Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index' with more than 40% null values

- Removing binary value columns like 'Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations' that have very high 'Yes-No' imbalance.

Missing Value and Outlier Treatment for Numerical Columns

- We see that columns TotalVisits and Page Views Per Visit have many outliers in the upper range.

- So, capped the two variables between 0.1 to 0.99 percentile.

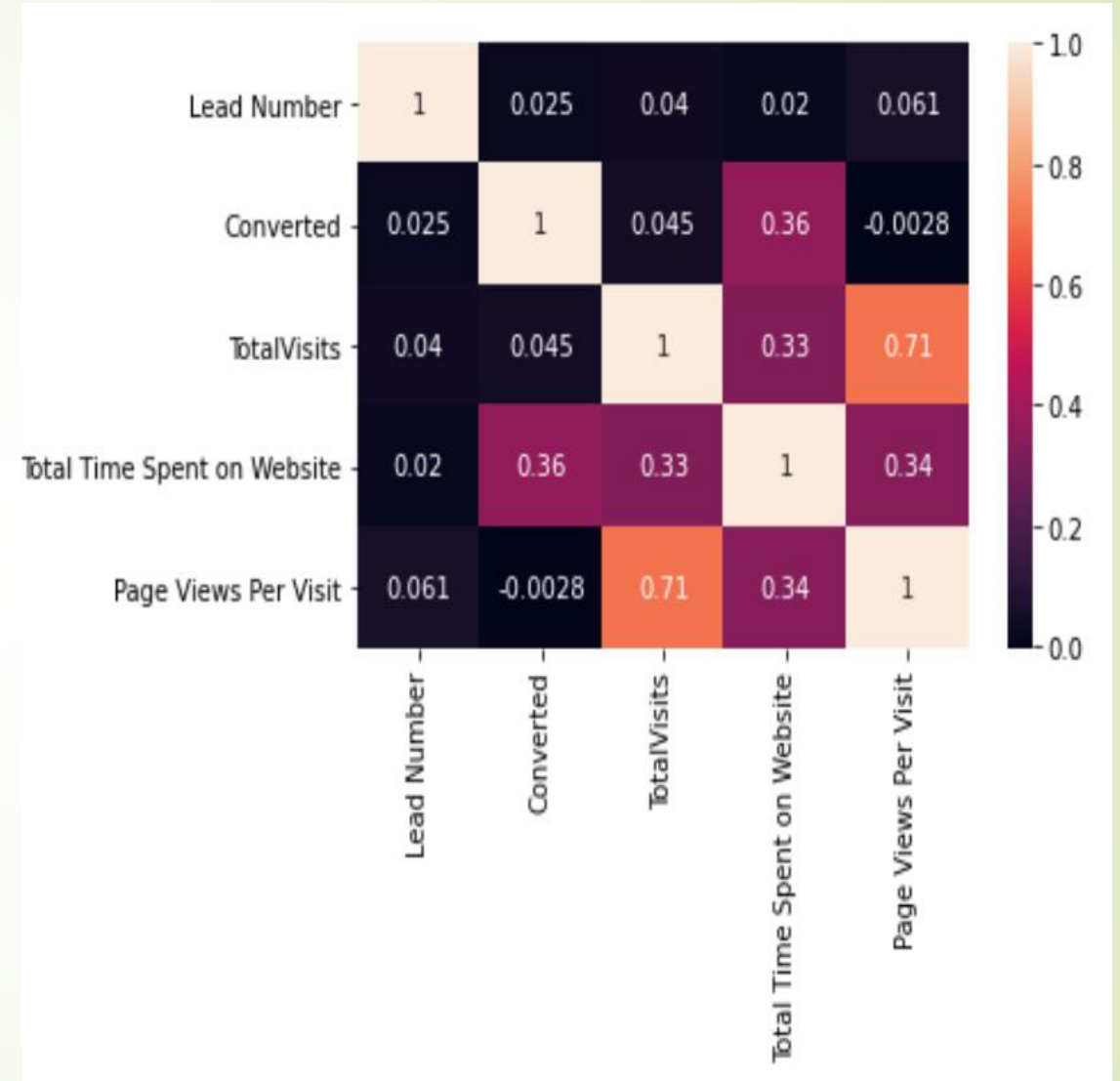- Imputing missing values in TotalVisits and Page Views Per Visit equal to mean
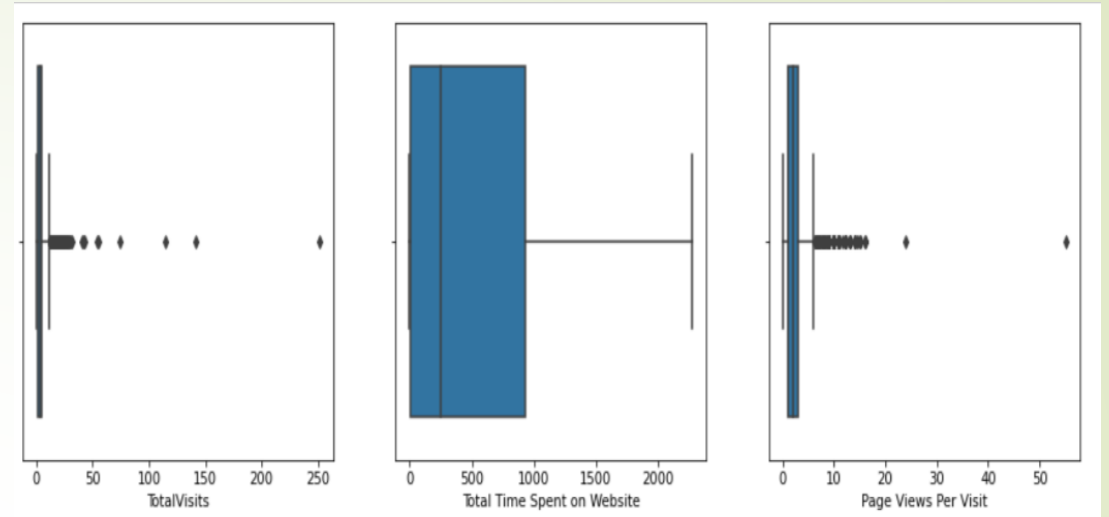
Before Outlier Treatment
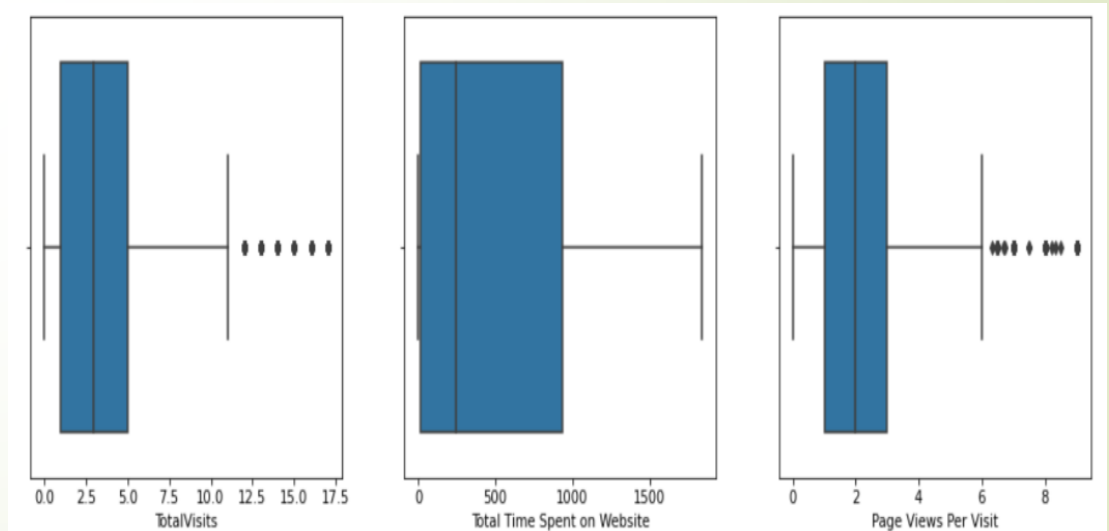
After Outlier Treatment

Multicollinearity check for Numerical columns

- Dropping 'Page Views Per Visit' since it has high correlation with 'TotalVisits'

## Missing value Treatment for Categorical Columns

- Since we will be dealing with the data that comes right from the forms filled by the people we are going to drop 'Tags' and 'Last Notable Activity', 'Last Activity' as it has been added by the sales team.

- Functions performed on Categorical columns:

  o Replacing similar values with one value

  o Replacing values with single digit count with value 'Other'

  o Replacing 'NaN' values with mode if percentage of missing values is low

  o Replacing 'NaN' values with 'Unknown' if percentage of missing values is high

- Final Categorical Columns:

  o 'Lead Origin'

  o 'Do Not Email',

  o 'A free copy of Mastering The Interview'

  o 'Country'

  o 'What is your current occupation'

  o 'What matters most to you in choosing a course'

  o 'Specialization'
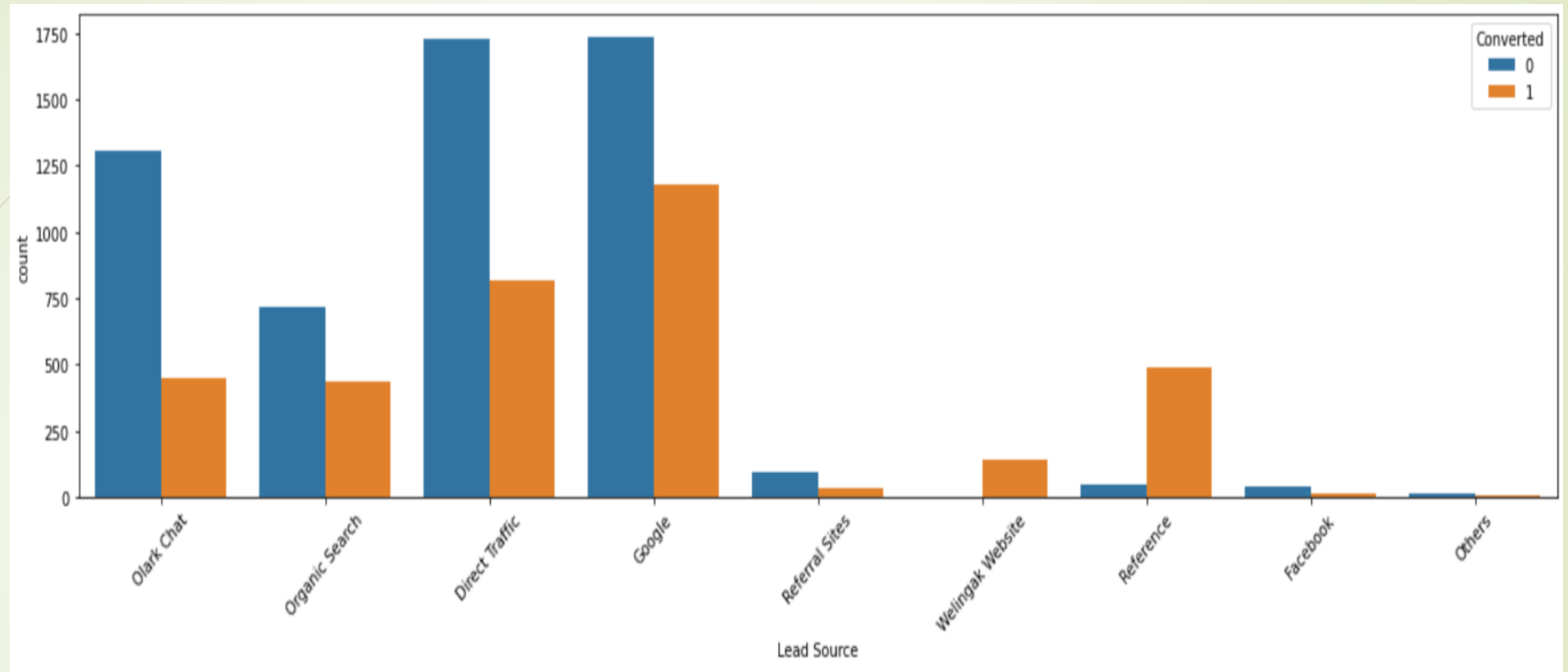
  o 'Lead Source'

Distribution of column Lead Source
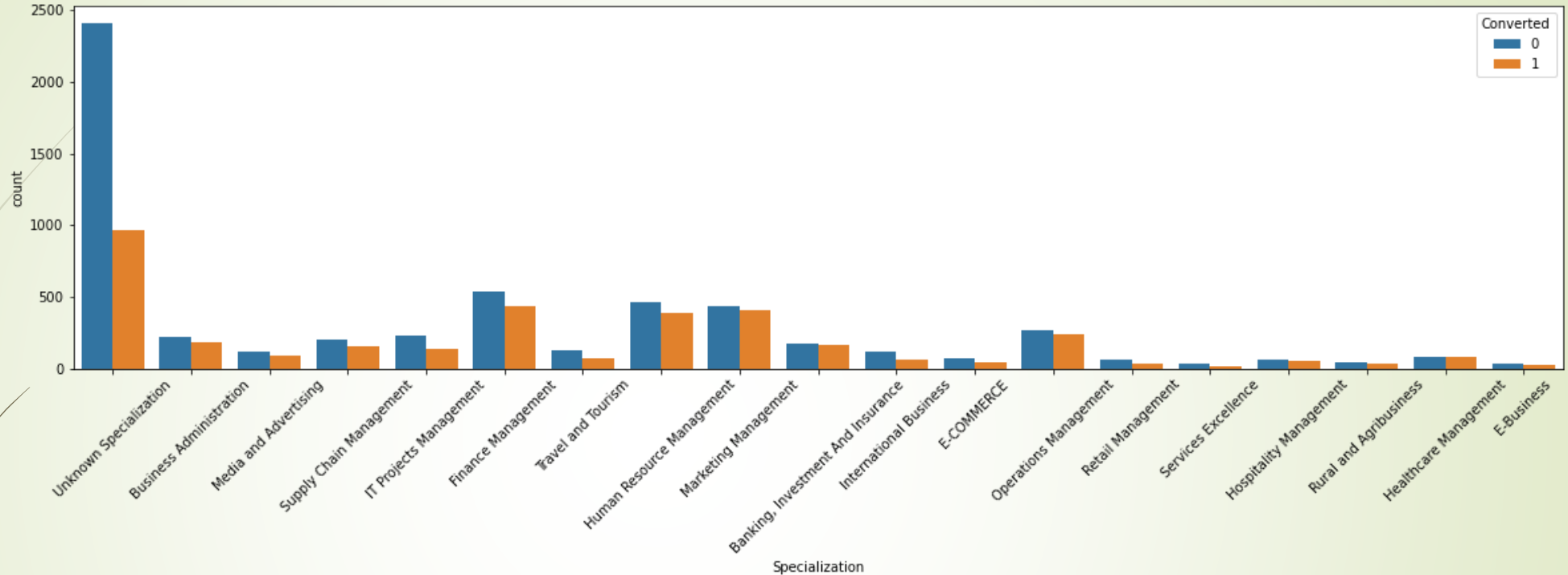
Inferences :

- 'Google', 'Olark Chat' and 'Direct Traffic' are the top converted categories in Lead Source.

- We should focus on Reference category as conversion changes are very high.

- We can see that for 'Welingak Website' all the leads have converted.

- 'Facebook' has high conversion rate
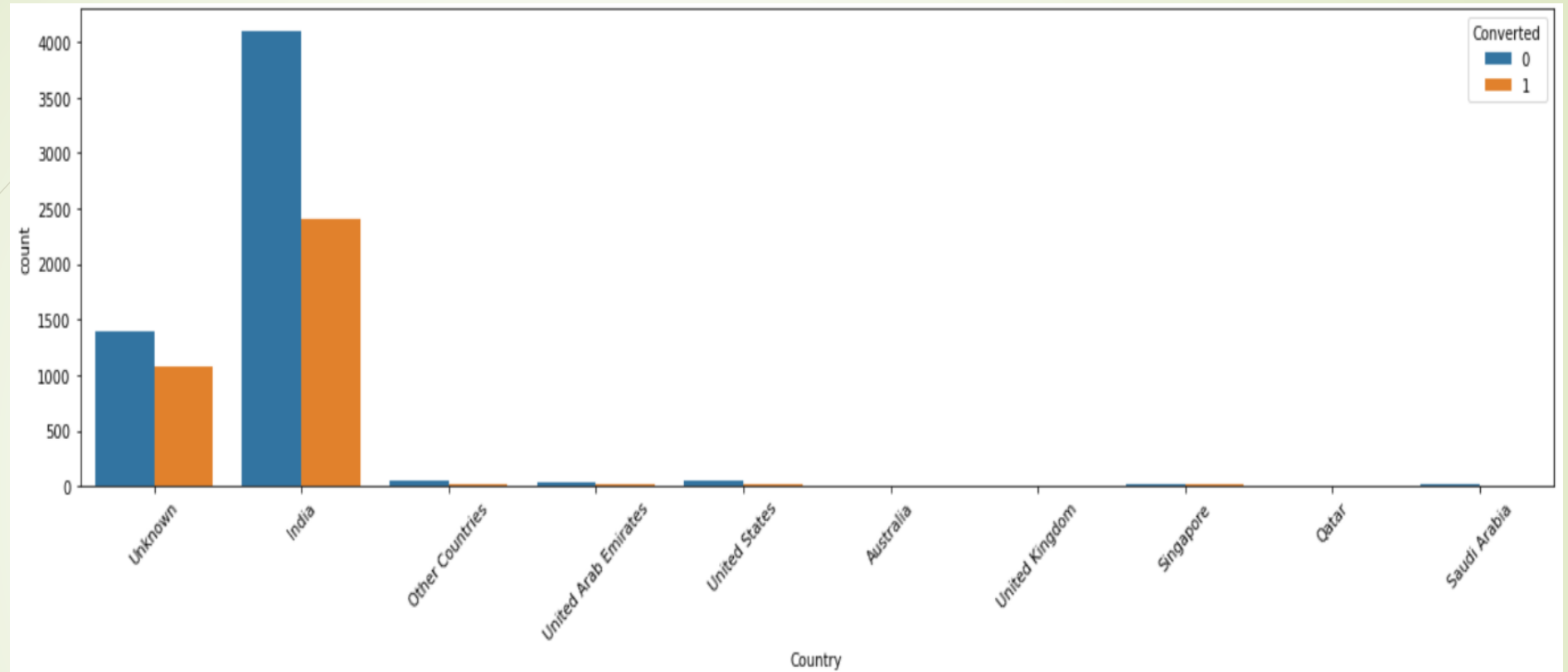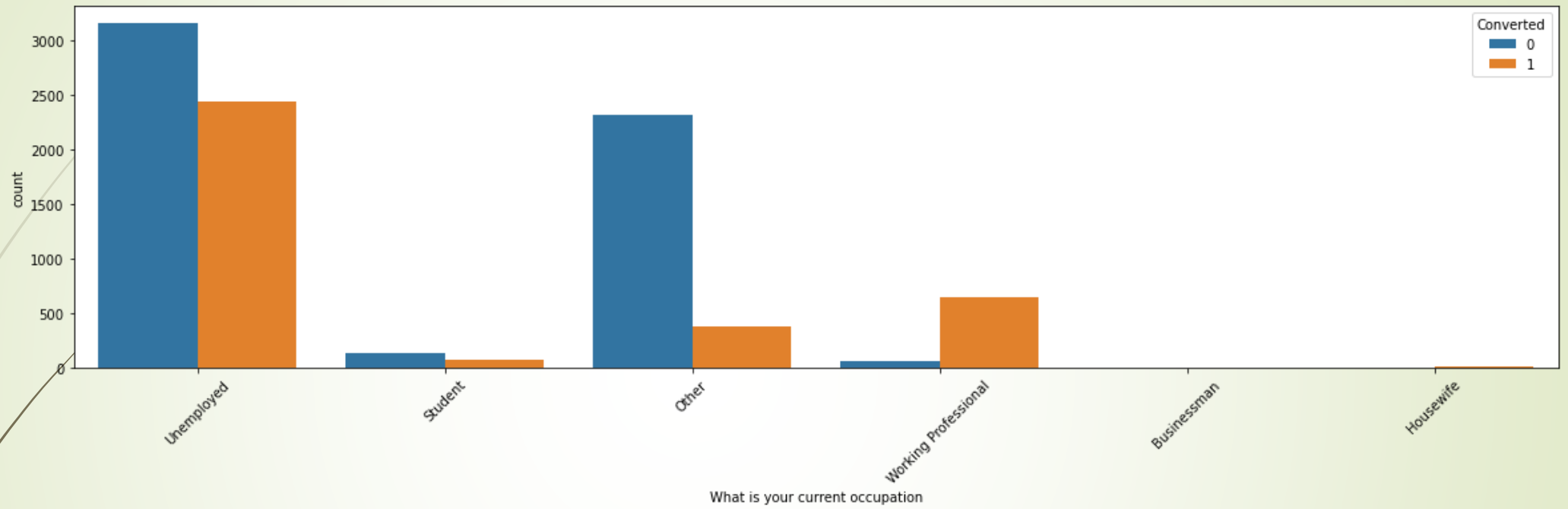
Distribution of Countries

Inference:

- We can see that most of the leads are from 'India' and number of conversions is also high.

# Distribution of column Specialization

Inference:

- Most of the specializations have higher conversion rate except 'Unknown'

- Top professionals applying for the course are HR, Finance, Marketing and Operations Management.

Distribution for the column What's your current Occupation?
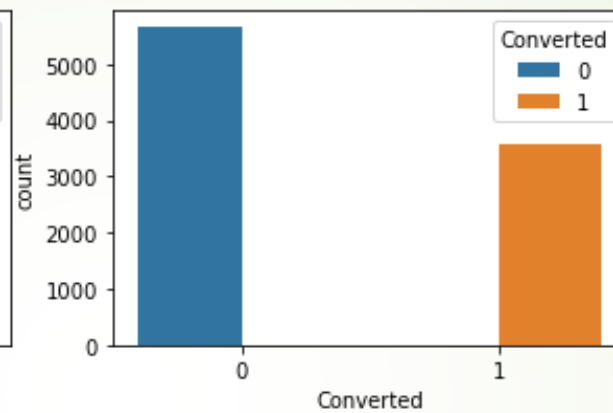
Inference:

- Working Professional and Housewife are more desirable to get converted .
- More number of unemployed people are applying and have better conversion.
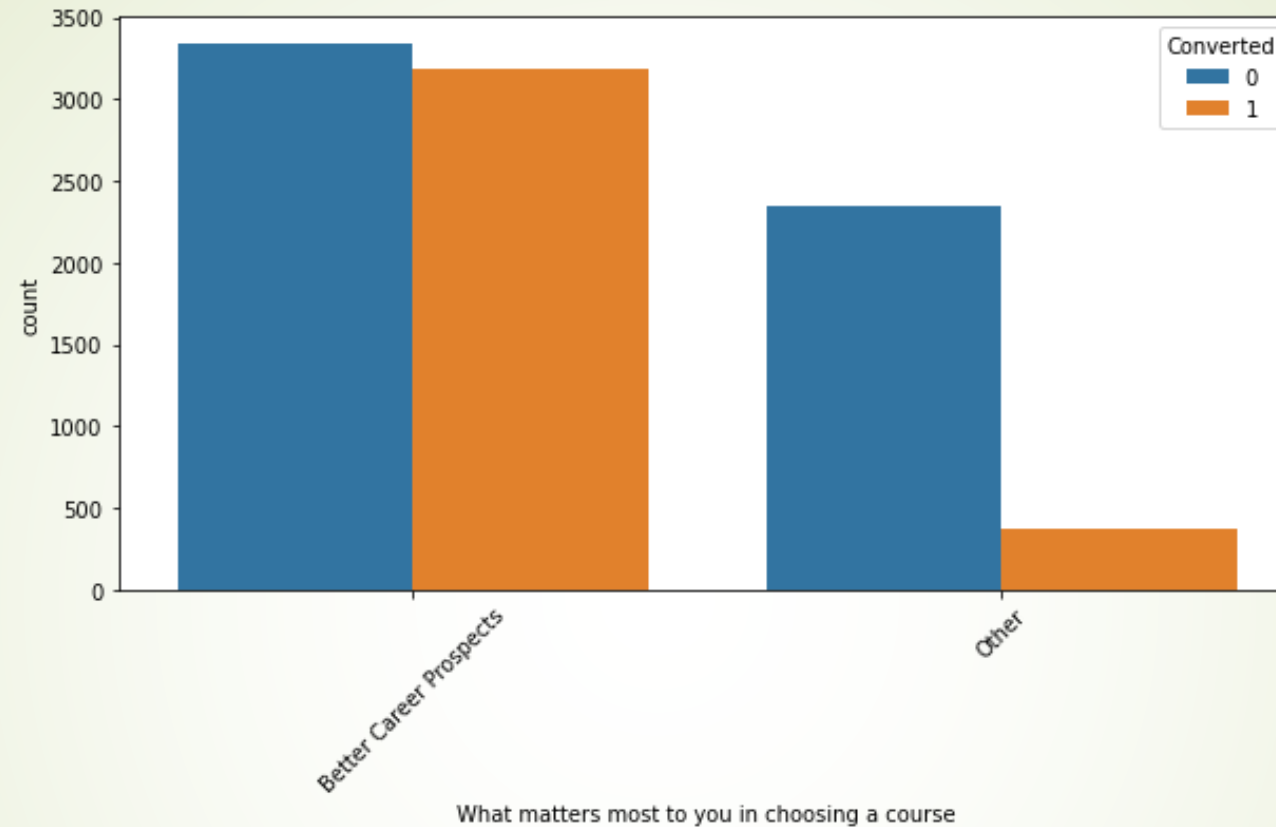
Distribution for column What matters most to you in choosing the course?

Inference:

• People looking for better career Prospects has more chances to get converted.
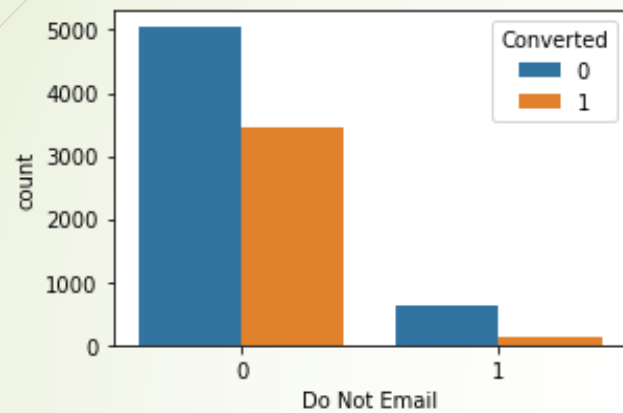
Numerical column vs Target variable

Inference:

• People who said 'No' to Do not Email have higher conversion rate.

Model Building

- After dropping unnecessary columns and creating the dummy variables we were left with 51 columns.

- For this dataset the imbalance percentage is 38.

- We have divided the dataset with 70:30 training and test split.

- We have run RFE with 15 columns and after 7 iterations we were left with 9 columns.

- We have plotted ROC curve with the training dataset which shows better performance.

Receiver operating characteristic example

Accuracy, Sensitivity and Specificity for various probabilities
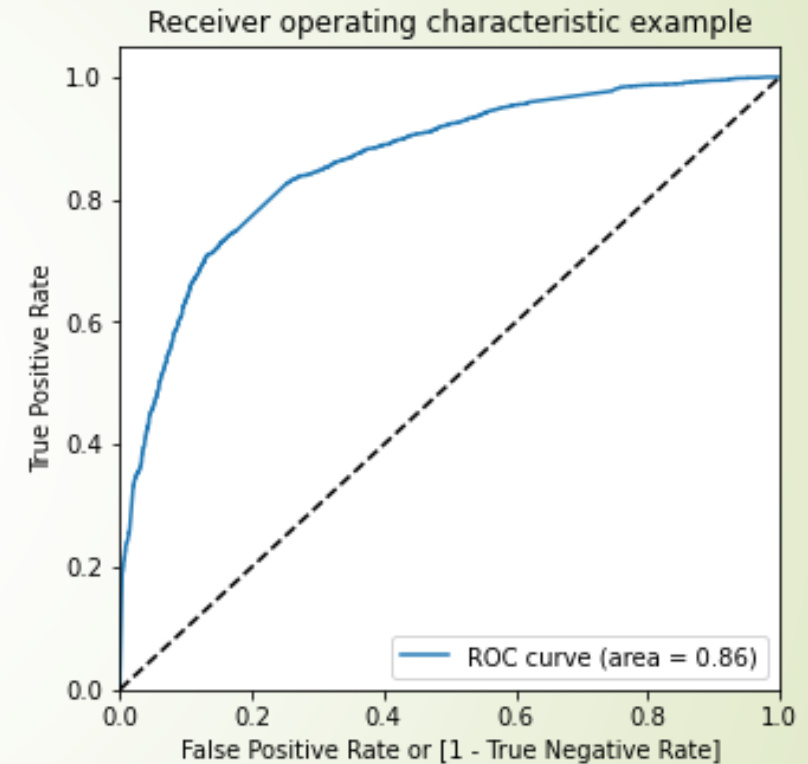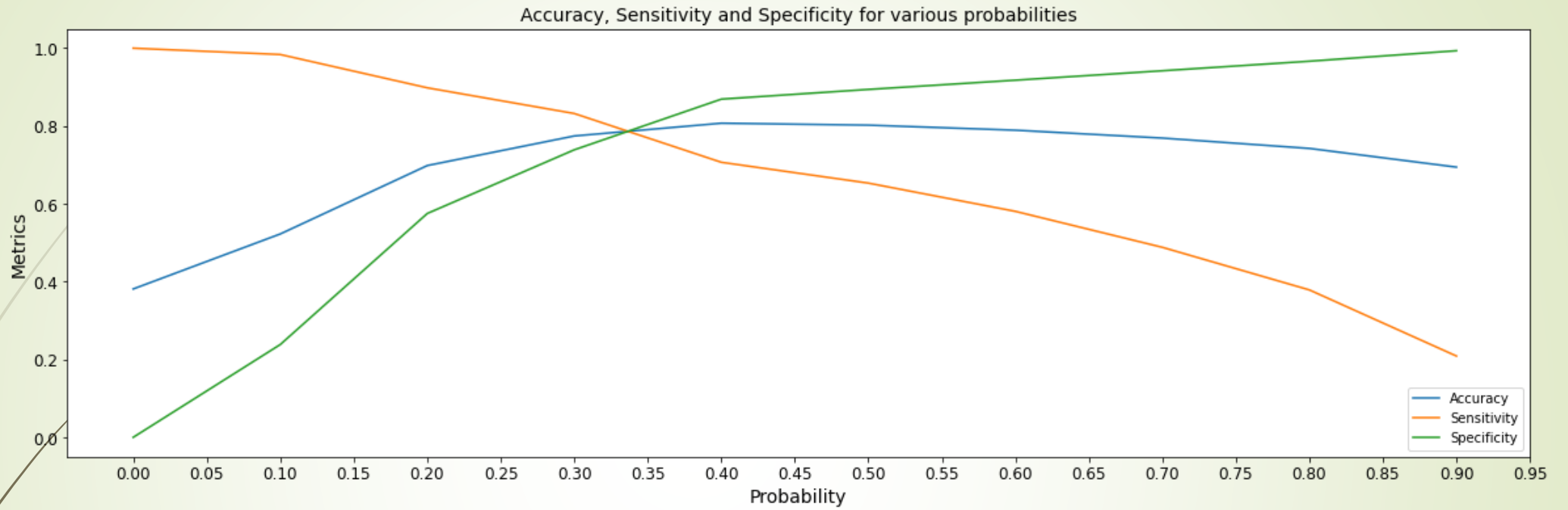
Finding Optimal Probability cut-off

From this curve we have found that the optimal probability cut-off is 0.34. So, for our prediction we have used cut-off Lead Score of 34.

Prediction metrics for Training dataset

	prob	Accuracy	Sensitivity	Specificity
0.0	0.0	0.381262	1.000000	0.000000
0.1	0.1	0.522573	0.983779	0.238381
0.2	0.2	0.698361	0.898216	0.575212
0.3	0.3	0.774428	0.832117	0.738881
0.4	0.4	0.807205	0.706813	0.869065
0.5	0.5	0.802257	0.653285	0.894053
0.6	0.6	0.789270	0.580697	0.917791
0.7	0.7	0.769017	0.488240	0.942029
0.8	0.8	0.742424	0.378751	0.966517
0.9	0.9	0.694341	0.208840	0.993503

{'accuracy': 79.96289424860854,
 'TN': 3357,
 'FP': 645,
 'FN': 651,
 'TP': 1815,
 'sensitivity': 73.60097323600974,
 'specificity': 83.88305847076461,
 'precision': 73.78048780487805,
 'False_Positive_Rate': 16.11694152923539,
 'Positve_Prediction_value': 0.7378048780487805,
 'Negative_Prediction_value': 0.8375748502994012}

Note: Lead Score=prob*100

Using a cut-off of 0.34 gives us accuracy of 80% and sensitivity of 73% which is not exactly matching with expectations to get conversion rate of 80%. If we decrease the probability cutoff then sensitivity will improve but resulting into decrease in our model accuracy. So as per business requirement we can adjust the probability cutoff.

Train dataset predictions using cut-off 0.31 (as per Business Requirement)

```
{'accuracy': 77.72108843537416,
 'TN': 2990,
 'FP': 1012,
 'FN': 429,
 'TP': 2037,
 'sensitivity': 82.60340632603406,
 'specificity': 74.71264367816092,
 'precision': 66.80878976713677,
 'False_Positive_Rate': 25.287356321839084,
 'Positve_Prediction_value': 0.6680878976713677,
 'Negative_Prediction_value': 0.8745247148288974}
```

We can see above that for our business requirement of lead conversion rate to be above 80%, we are getting a model which with 77% accuracy gives us a sensitivity of 82%

Decision Making

- Now, if we use cut-ff of 0.31 we will correctly identify all the leads who will convert

- But, we will also incorrectly identify some leads as 'will convert' which will not convert.

- This can use some extra time.

- As the business problem does not state any hard deadline to meet a certain target for the Sales Team, we can go with 0.31 cut-off or Lead Score=31

Prediction metrics for Test dataset

```
{'accuracy': 78.10245310245311,
 'TN': 1245,
 'FP': 432,
 'FN': 175,
 'TP': 920,
 'sensitivity': 84.01826484018264,
 'specificity': 74.23971377459749,
 'precision': 68.04733727810651,
 'False_Positive_Rate': 25.76028622540251,
 'Positve_Prediction_value': 0.6804733727810651,
 'Negative_Prediction_value': 0.8767605633802817}
```

We can see the accuracy of test data set is 78% and sensitivity is 84% which is matching with Training dataset metrics. So our model is performing good on both training and test dataset.

The Final Equation

```
Lead Origin_Lead Add Form            3.735612
Total Time Spent on Website          3.690194
Occupation_Working Professional      2.568117
Lead Source_Welingak Website         2.113739
Matters_Better Career Prospects      1.266959
Lead Source_Olark Chat               1.042105
TotalVisits                          0.854924
Lead Origin_API                      0.174739
Do Not Email                        -1.299315
```

Odds(**Converted**)='(3.74)***Lead Origin_Lead Add Form** + (3.69)***Total Time Spent on Website** + (2.57)***Occupation_Working Professional** + (2.11)***Lead Source_Welingak Website** + (1.27)***Matters_Better Career Prospects** + (1.04)***Lead Source_Olark Chat** + (0.85)***TotalVisits** + (0.17)***Lead Origin_API** + (-1.3)***Do Not Email** + (-3.27)'
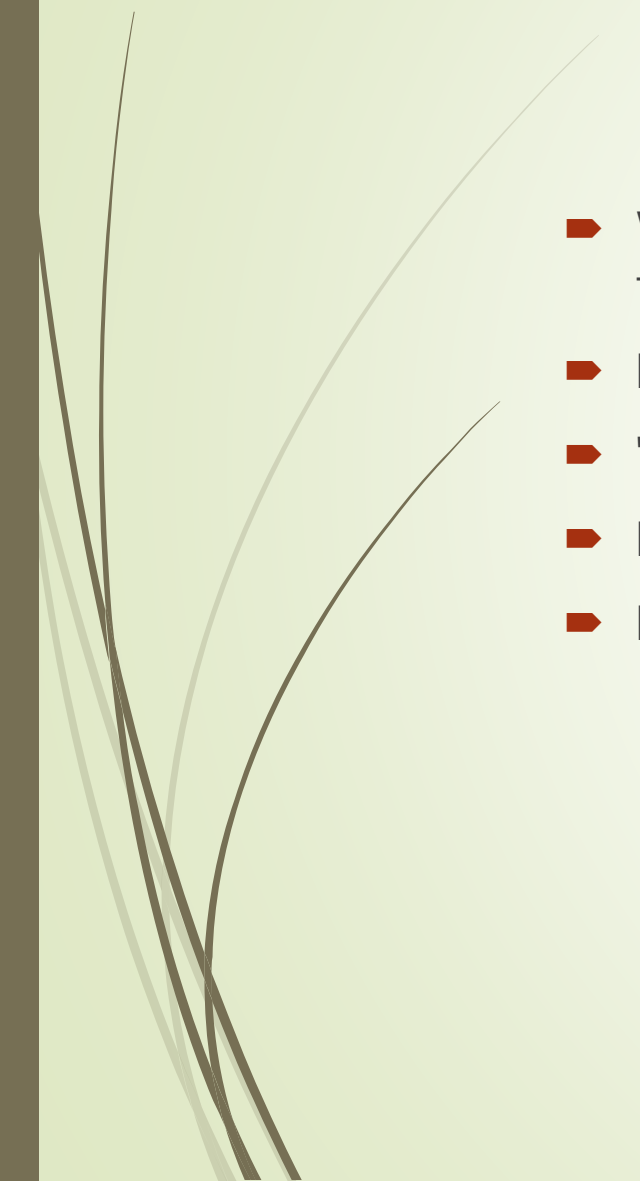
Conclusion

- X Education has to focus on following variables to identify Hot Leads:
  - Total Time Spent on Website
  - Lead Origin_Lead Add Form
  - Occupation_Working Professional
  - Lead Source_Welingak Website
  - Do Not Email
  - Matters_Better Career Prospects
  - Lead Source_Olark Chat
  - TotalVisits
  - Lead Origin_API
- This model is adjustable to the company's requirement . We can change the Lead Score cut off to change the sensitivity and accuracy of the model. This will depend on business requirement.

Inferences from EDA

- We can see that if the lead source is 'Reference' and 'Welingak Website' then there is high chances of conversion.

- People who sent 'SMS' have high chances of conversion

- 'Working Professional' and 'Housewife' are more desirable to get converted

- People looking for 'Better Career Prospects' has more chances to convert

- People who said No to 'Do Not Email' has higher conversion chance