## Case study objective

An education company named X Education sells online courses to industry professionals.

- The objective of this case study is to select the most promising leads, i.e., the leads that are most likely to convert into paying customers, also known as “hot leads”
- Build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance
- Target is to obtain lead conversion rate to be around 80%.

## Problem Solving Methodology

### Phase 1: Business Understanding

Identify the most potential leads, also known as “hot leads,” to increase the conversion rate.

### Phase 2: Data Understanding

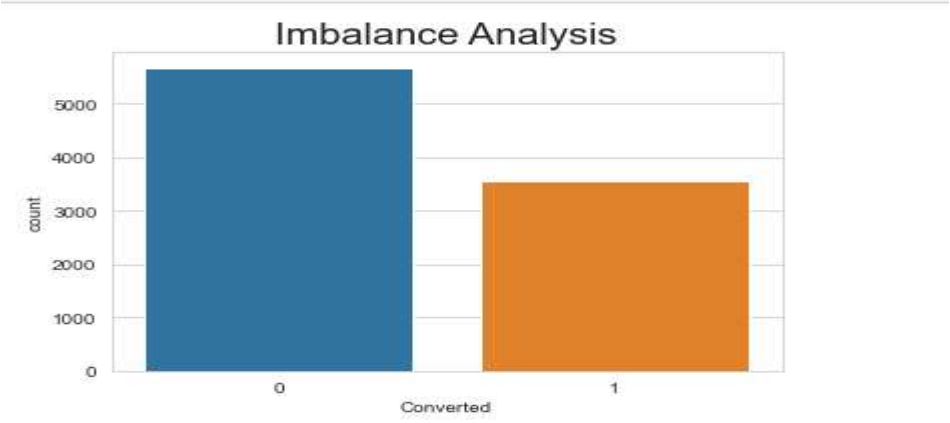
- Dataset provided for analysis is the Leads dataset from the past with around 9000 data points.
- The dataset consists of various attributes such as Lead Source, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not, wherein 1 means it was converted and 0 means it wasn’t converted.

### Phase 3: Data Preparation

- Identification of important variables
  a. Target variable and feature variables
  b. Data types
  c. Categorical or Continuous variable

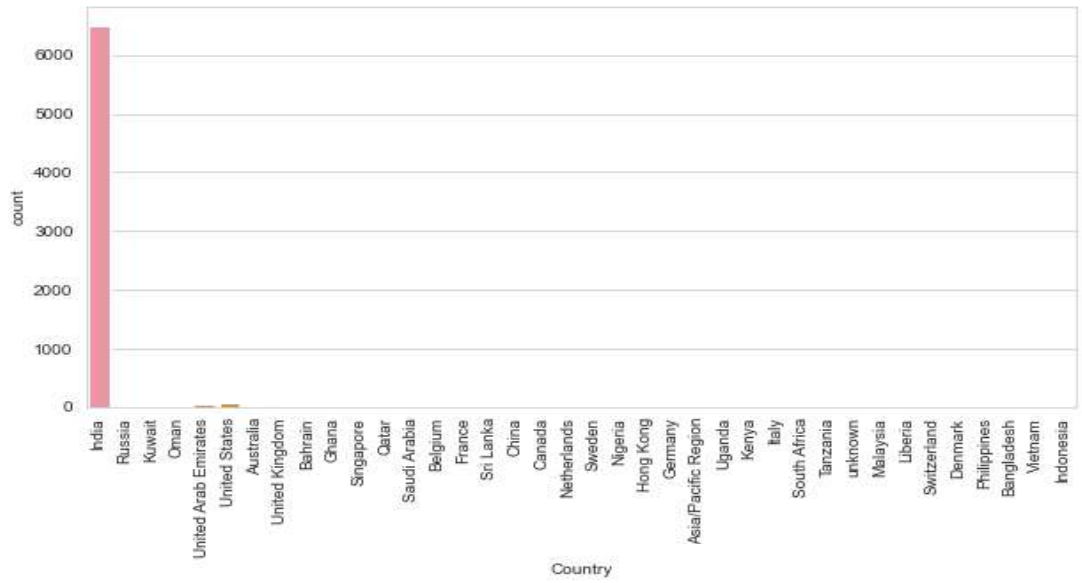
- Handled “Select” level in multiple categorical columns by replacing it with null value.
- Handled missing values by deleting columns which has more than 40% of missing values.
- Checked outliers in numerical columns and performed outlier treatment.
- Identified null values in all columns and replaced them with appropriate values.
- Converted binary variables Yes/No to 1/0.
- Dropped the columns generated by the Sales team.
- Checked the data imbalance and the imbalance ratio.
- Combined levels of a categorical variable having fewer value counts.
- Visualized data with respect to target variable.
- Performed univariate, bivariate, and multivariate analysis.
- Created dummy variables for categorical variables with multiple levels.
- Performed 70%-30% train-test split.
- Identified and performed scaling of numerical variables using Standard Scalar.

**Imbalance analysis**

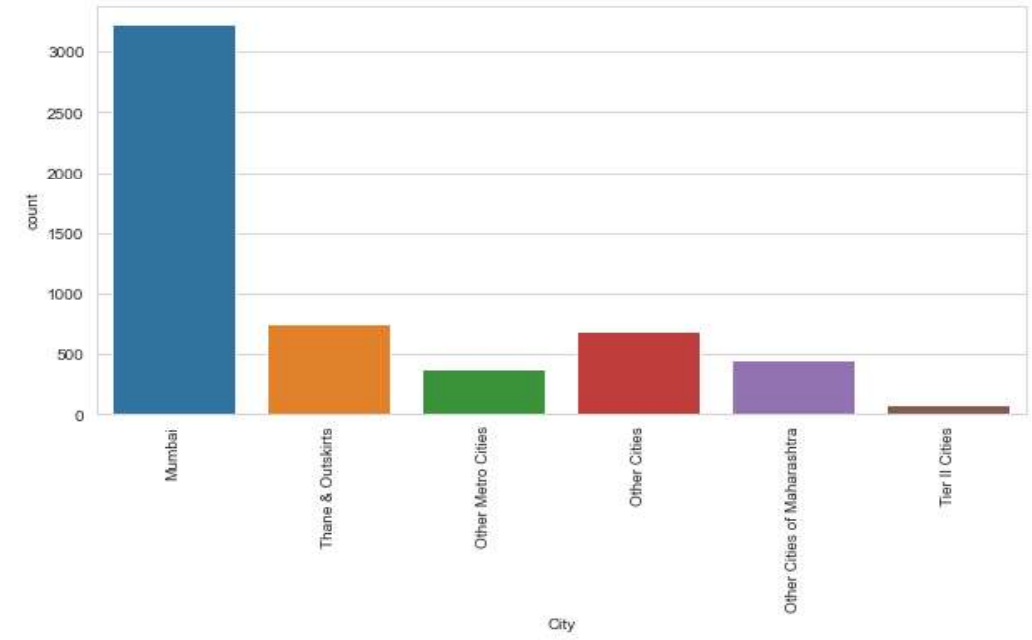



Conversion rate is 37% and the imbalance ratio is 1.59

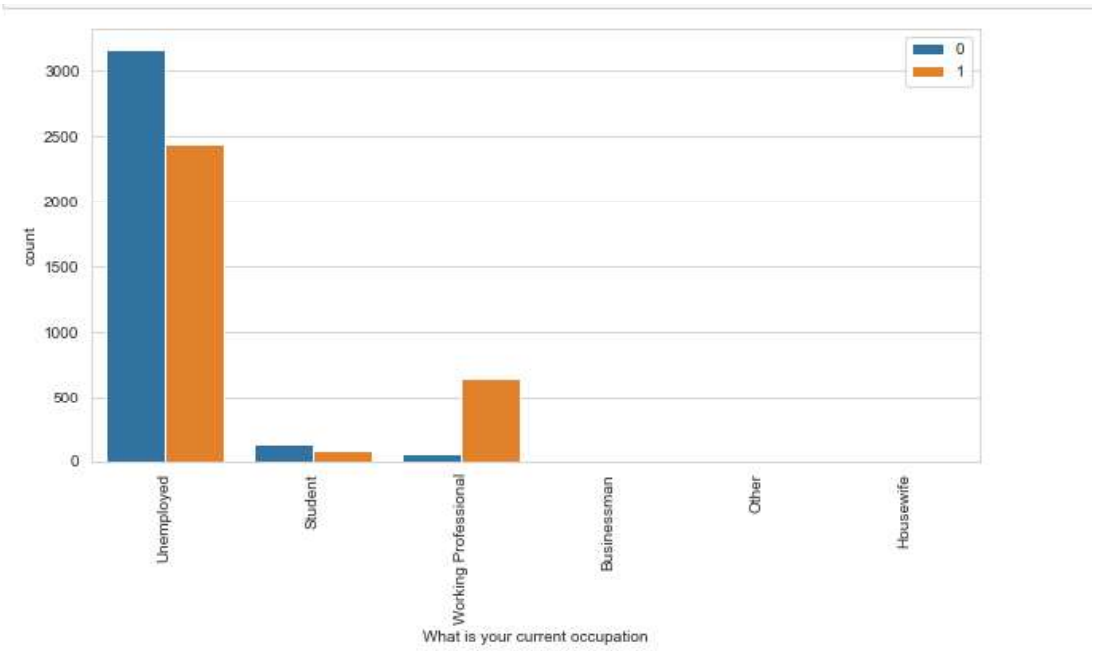
## Univariate and Bivariate Analysis



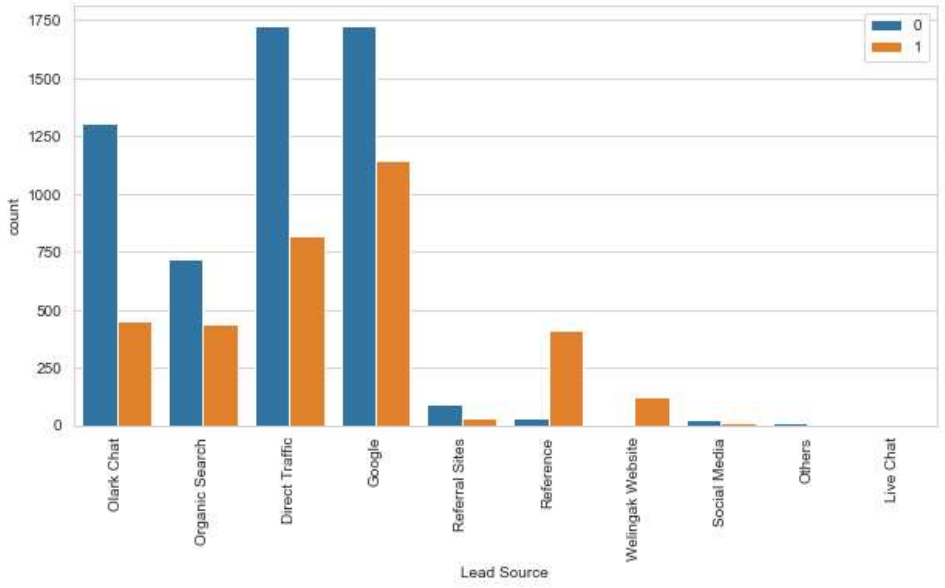

- More than 90% of the value is India in Country column.



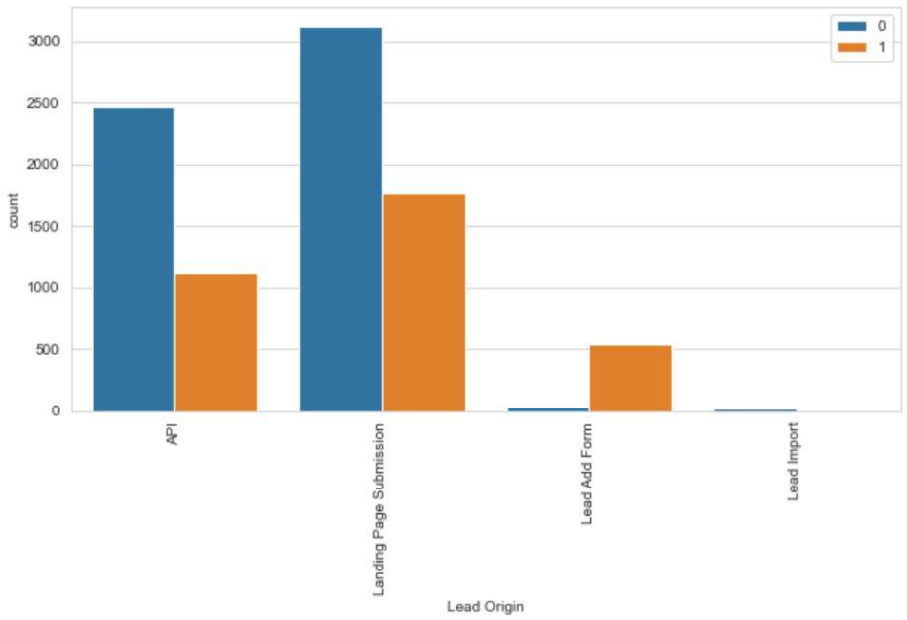

- The column cannot be imputed with mode as it makes the whole data skewed.
- Also, X-Education is online teaching platform. The city and country information will not be much useful as potential students can available any courses online despite their city. Hence, country and city columns are not required for Model building.



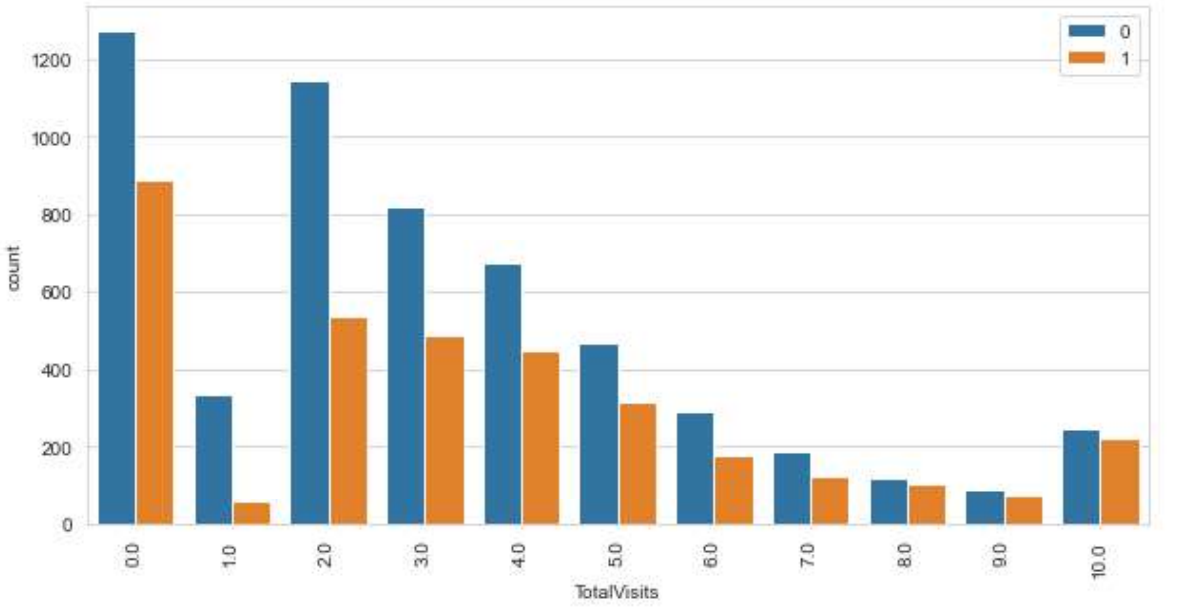

- The Conversion rate is high for Working Professionals. It can be an important column for model building.
- Unemployed has high conversion rate. Count of Unemployed is high in the dataset.

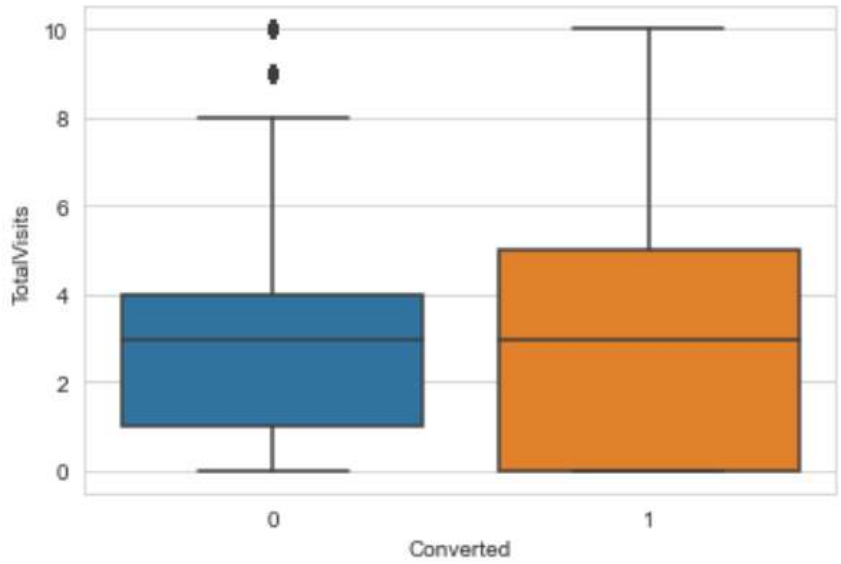



- Majority traffic is from Google and Direct traffic
- Conversion Rate of reference leads and leads through Welingak website is high.

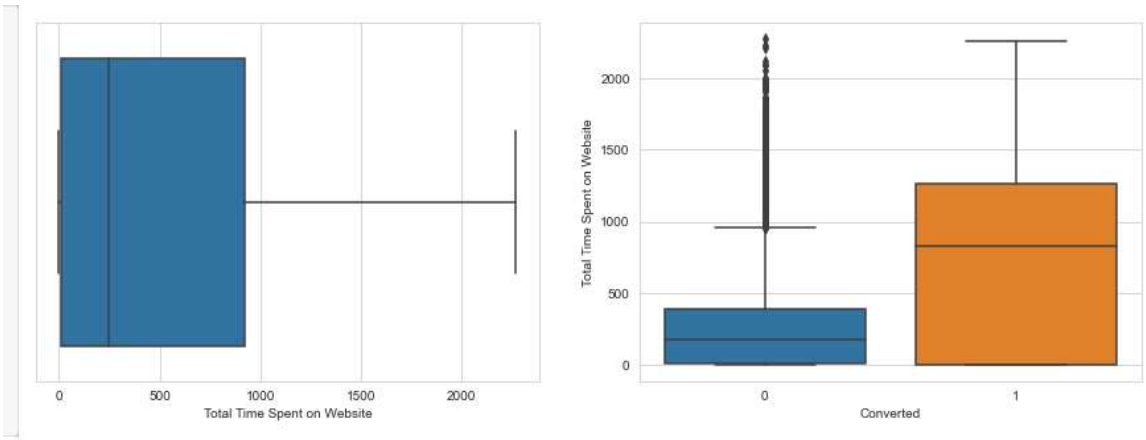



- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import are very less in count and conversion rate is also the lowest
- To improve overall lead conversion rate, we must improve lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.






- Median for converted and not converted leads are the same.
- Higher conversion rate with increase in the total number of visits per customer.




- Customers who spend more time on the website are more likely to be converted.

## Summary of feature and target variables

```
The Target Feature is : ['Converted']

The Boolean Features are : ['Do Not Email', 'A free copy of Mastering The Interview']

The Categorical Features are : ['Lead Origin', 'Lead Source', 'Specialization', 'What is your current occupation']

The Numeric Features are : ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']
```

## Correlation of variables

```
#finding Top 10 Positively Correlated values

corrdf.sort_values(by='Correlation',ascending=False).head(10)
```

| | Var1 | Var2 | Correlation |
|---|---|---|---|
| 504 | Lead Source_Social Media | Lead Origin_Lead Import | 0.95 |
| 441 | Lead Source_Reference | Lead Origin_Lead Add Form | 0.87 |
| 126 | Page Views Per Visit | TotalVisits | 0.77 |
| 191 | Lead Origin_Landing Page Submission | A free copy of Mastering The Interview | 0.56 |
| 190 | Lead Origin_Landing Page Submission | Page Views Per Visit | 0.55 |
| 879 | Specialization_Not Specified | Lead Source_Olark Chat | 0.51 |
| 812 | Specialization_Management_Specializations | Lead Origin_Landing Page Submission | 0.50 |
| 534 | Lead Source_Welingak Website | Lead Origin_Lead Add Form | 0.46 |
| 188 | Lead Origin_Landing Page Submission | TotalVisits | 0.45 |
| 94 | Total Time Spent on Website | Converted | 0.36 |

```
#finding Top 10 Negatively  Correlated values

corrdf.sort_values(by='Correlation',ascending=True).head(10)
```

| | Var1 | Var2 | Correlation |
|---|---|---|---|
| 703 | What is your current occupation_Unknown | What is your current occupation_Unemployed | -0.80 |
| 874 | Specialization_Not Specified | Lead Origin_Landing Page Submission | -0.76 |
| 894 | Specialization_Not Specified | Specialization_Management_Specializations | -0.70 |
| 345 | Lead Source_Olark Chat | Page Views Per Visit | -0.58 |
| 347 | Lead Source_Olark Chat | Lead Origin_Landing Page Submission | -0.53 |
| 873 | Specialization_Not Specified | A free copy of Mastering The Interview | -0.51 |
| 343 | Lead Source_Olark Chat | TotalVisits | -0.50 |
| 872 | Specialization_Not Specified | Page Views Per Visit | -0.46 |
| 870 | Specialization_Not Specified | TotalVisits | -0.40 |
| 344 | Lead Source_Olark Chat | Total Time Spent on Website | -0.38 |

## Phase 4: Model Building

- Recursive Feature Elimination (RFE) was used to select top 20 variables.

  Top 20 Feature selected by RFE are

```
Index(['Do Not Email', 'TotalVisits', 'Total Time Spent on Website',
       'A free copy of Mastering The Interview',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Origin_Lead Import', 'Lead Source_Live Chat',
       'Lead Source_Olark Chat', 'Lead Source_Others', 'Lead Source_Reference',
       'Lead Source_Referral Sites', 'Lead Source_Social Media',
       'Lead Source_Welingak Website',
       'What is your current occupation_Housewife',
       'What is your current occupation_Other',
       'What is your current occupation_Unknown',
       'What is your current occupation_Working Professional',
       'Specialization_Not Specified', 'Specialization_Travel and Tourism'],
      dtype='object')
```

- Logistic Regression Model was built and VIF and p-values were checked. Model built in the first iteration has 20 variables

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6351 |
| Model: | GLM | Df Residuals: | 6330 |
| Model Family: | Binomial | Df Model: | 20 |
| Link Function: | logit | Scale: | 1.0000 |
| Method: | IRLS | Log Likelihood: | 2798.4 |
| Date: | Sat, 12 Jun 2021 | Deviance: | 5596.8 |
| Time: | 21:37:05 | Pearson chi2: | 6.70e+03 |
| No. Iterations: | 21 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.0830 | 0.120 | -0.690 | 0.490 | -0.318 | 0.153 |
| Do Not Email | -1.3440 | 0.169 | -7.966 | 0.000 | -1.675 | -1.013 |
| TotalVisits | 0.1949 | 0.043 | 4.569 | 0.000 | 0.111 | 0.279 |
| Total Time Spent on Website | 1.0742 | 0.039 | 27.527 | 0.000 | 0.998 | 1.151 |
| A free copy of Mastering The Interview | -0.2690 | 0.086 | -3.116 | 0.002 | -0.438 | -0.100 |
| Lead Origin_Landing Page Submission | -0.6325 | 0.126 | -5.036 | 0.000 | -0.879 | -0.386 |
| Lead Origin_Lead Add Form | 0.9265 | 1.079 | 0.859 | 0.390 | -1.188 | 3.041 |
| Lead Origin_Lead Import | 21.9419 | 3.4e+04 | 0.001 | 0.999 | -6.67e+04 | 6.68e+04 |
| Lead Source_Live Chat | 23.8011 | 3.41e+04 | 0.001 | 0.999 | -6.68e+04 | 6.68e+04 |
| Lead Source_Olark Chat | 1.2607 | 0.128 | 9.877 | 0.000 | 1.010 | 1.511 |
| Lead Source_Others | 0.2381 | 0.799 | 0.298 | 0.766 | -1.328 | 1.804 |
| Lead Source_Reference | 2.7967 | 1.099 | 2.545 | 0.011 | 0.643 | 4.950 |
| Lead Source_Referral Sites | -0.4787 | 0.328 | -1.459 | 0.145 | -1.122 | 0.165 |
| Lead Source_Social Media | -20.7550 | 3.4e+04 | -0.001 | 1.000 | -6.68e+04 | 6.67e+04 |
| Lead Source_Welingak Website | 5.1201 | 1.298 | 3.944 | 0.000 | 2.576 | 7.664 |
| What is your current occupation_Housewife | 21.3080 | 2.14e+04 | 0.001 | 0.999 | -4.2e+04 | 4.2e+04 |
| What is your current occupation_Other | 0.3914 | 0.720 | 0.543 | 0.587 | 1.803 | 1.020 |
| What is your current occupation_Unknown | -1.2943 | 0.086 | -15.086 | 0.000 | -1.463 | -1.126 |
| What is your current occupation_Working Professional | 2.4025 | 0.182 | 13.201 | 0.000 | 2.046 | 2.759 |
| Specialization_Not Specified | -0.9083 | 0.119 | -7.626 | 0.000 | -1.142 | -0.675 |
| Specialization_Travel and Tourism | -0.2581 | 0.237 | -1.090 | 0.276 | -0.722 | 0.206 |

- Using Manual Feature Elimination, a variable was recursively eliminated by checking the VIF and p-values. This is repeated until VIF and p-values in the model are significant.

Final model has 11 variables, p-value of all variables is <0.05 and VIF is less than 3.

Generalized Linear Model Regression Results

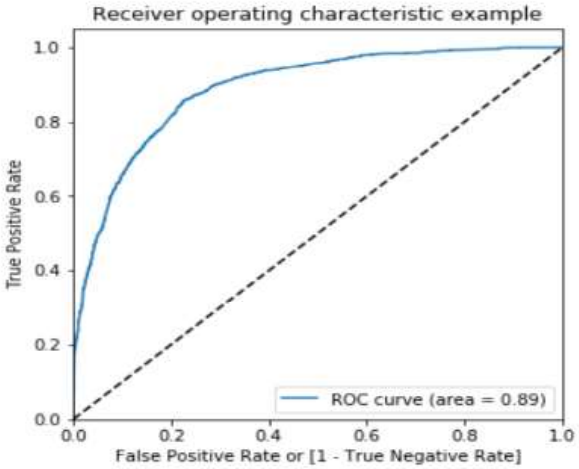
| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6351 |
| Model: | GLM | Df Residuals: | 6339 |
| Model Family: | Binomial | Df Model: | 11 |
| Link Function: | logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -2807.5 |
| Date: | Sat, 12 Jun 2021 | Deviance: | 5615.0 |
| Time: | 21:37:07 | Pearson chi2: | 6.70e+03 |
| No. Iterations: | 7 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.0612 | 0.118 | -0.520 | 0.603 | -0.292 | 0.170 |
| Do Not Email | -1.3548 | 0.168 | -8.067 | 0.000 | -1.684 | -1.026 |
| TotalVisits | 0.1759 | 0.042 | 4.172 | 0.000 | 0.093 | 0.259 |
| Total Time Spent on Website | 1.0656 | 0.039 | 27.515 | 0.000 | 0.990 | 1.141 |
| A free copy of Mastering The Interview | -0.2788 | 0.086 | -3.247 | 0.001 | -0.447 | -0.111 |
| Lead Origin_Landing Page Submission | -0.6463 | 0.124 | -5.217 | 0.000 | -0.889 | -0.403 |
| Lead Origin_Lead Add Form | 3.6276 | 0.242 | 14.994 | 0.000 | 3.153 | 4.102 |
| Lead Source_Olark Chat | 1.2230 | 0.125 | 9.747 | 0.000 | 0.977 | 1.469 |
| Lead Source_Welingak Website | 2.3856 | 0.754 | 3.165 | 0.002 | 0.908 | 3.863 |
| What is your current occupation_Unknown | -1.2950 | 0.086 | -15.138 | 0.000 | -1.463 | -1.127 |
| What is your current occupation_Working Professional | 2.4028 | 0.182 | 13.218 | 0.000 | 2.046 | 2.759 |
| Specialization_Not Specified | -0.9232 | 0.118 | -7.805 | 0.000 | -1.155 | -0.691 |

| | Features | VIF |
|---|---|---|
| 4 | Lead Origin_Landing Page Submission | 2.58 |
| 6 | Lead Source_Olark Chat | 2.21 |
| 3 | A free copy of Mastering The Interview | 2.19 |
| 10 | Specialization_Not Specified | 2.04 |
| 1 | TotalVisits | 1.64 |
| 8 | What is your current occupation_Unknown | 1.58 |
| 5 | Lead Origin_Lead Add Form | 1.56 |
| 7 | Lead Source_Welingak Website | 1.33 |
| 2 | Total Time Spent on Website | 1.29 |
| 9 | What is your current occupation_Working Profes... | 1.16 |
| 0 | Do Not Email | 1.11 |

- Initially, an arbitrary cutoff value of 0.5 was used to calculate the predicted probably of getting converted on the train dataset.
- ROC curve: Trade-off between TPR and FPR. Higher the area under the curve, the better the model. Area = 0.89
- Accuracy of the model is 80.32%. Accuracy specifies the model performance based on both class (converted and not converted), while in our case study only a single class is important(converted). Hence, we check sensitivity, specificity, recall, precision, positive prediction value, negative prediction value, true positive rate, and false positive rate.

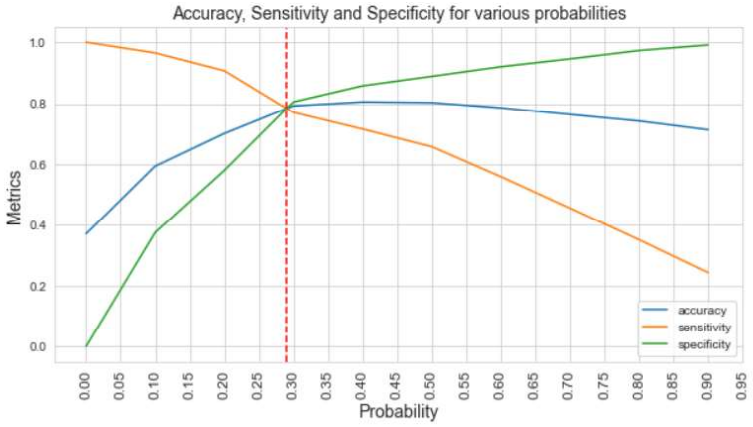
```
Model Accuracy value is             :  80.32 %
Model Sensitivity value is          :  65.79 %
Model Specificity value is          :  88.89 %
Model Precision value is            :  77.73 %
Model Recall value is               :  88.89 %
Model True Positive Rate (TPR)      :  65.79 %
Model False Positive Rate (FPR)     :  11.11 %
Model Poitive Prediction Value is   :  77.73 %
Model Negative Prediction value is  :  81.5 %
Model F1 score is                   :  82.94 %
```




- To improve sensitivity, the cut-off point/threshold is redefined.

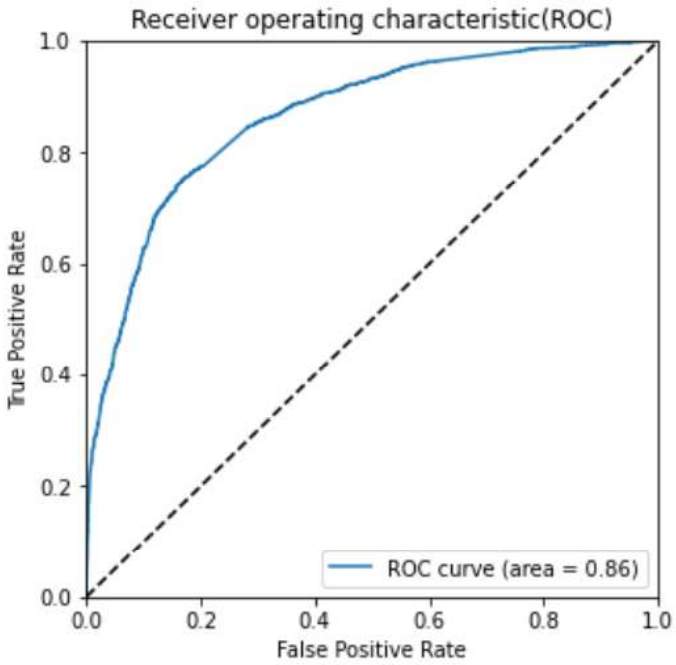
**Phase 5: Model evaluation:**

- Accuracy, Sensitivity, and Specificity is calculated at different cut off values and stored in a data frame.
- Optimal threshold point was found using Accuracy, Sensitivity, and Specificity tradeoff graph.



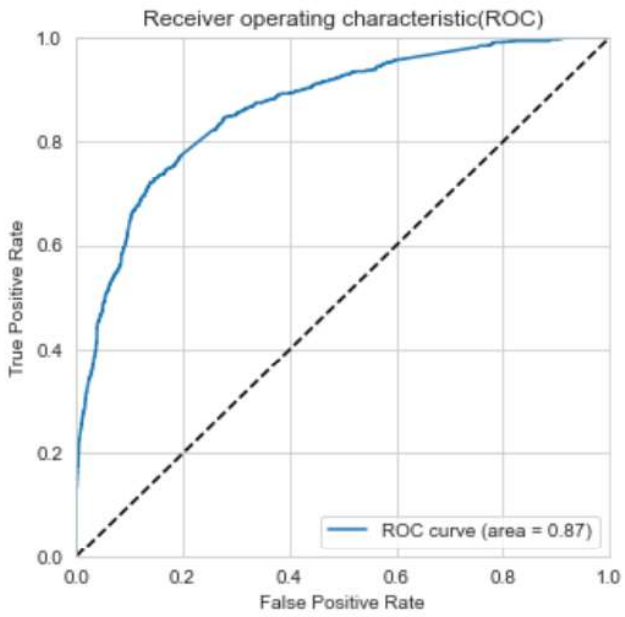
- The conversion probability was calculated using the optimal cutoff point and model's performance was checked over train dataset.

```
Model Accuracy value is              :  78.96 %
Model Sensitivity value is           :  77.55 %
Model Specificity value is           :  79.8 %
Model Precision value is             :  69.36 %
Model Recall value is                :  79.8 %
Model True Positive Rate (TPR)       :  77.55 %
Model False Positive Rate (FPR)      :  20.2 %
Model Poitive Prediction Value is    :  69.36 %
Model Negative Prediction value is   :  85.77 %
Model F1 score is                    :  74.22 %
```




- The model was tested using optimal cutoff point and the model performance was checked using test dataset.

```
Model Accuracy value is              :  79.18 %
Model Sensitivity value is           :  77.2 %
Model Specificity value is           :  80.47 %
Model Precision value is             :  72.18 %
Model Recall value is                :  80.47 %
Model True Positive Rate (TPR)       :  77.2 %
Model False Positive Rate (FPR)      :  19.53 %
Model Poitive Prediction Value is    :  72.18 %
Model Negative Prediction value is   :  84.32 %
Model F1 score is                    :  76.1 %
```




## Driver variables for lead scoring

| | |
|---|---|
| Lead Origin_Lead Add Form | 3.627629 |
| What is your current occupation_Working Professional | 2.402787 |
| Lead Source_Welingak Website | 2.385581 |
| Lead Source_Olark Chat | 1.222974 |
| Total Time Spent on Website | 1.065566 |
| TotalVisits | 0.175917 |
| const | -0.061248 |
| A free copy of Mastering The Interview | -0.278838 |
| Lead Origin_Landing Page Submission | -0.646276 |
| Specialization_Not Specified | -0.923186 |
| What is your current occupation_Unknown | -1.295032 |
| Do Not Email | -1.354841 |

The company should make calls to the leads who match the below criteria as these are more likely to get converted:

- lead sources such as Welingak Websites and Olark Chat.
- leads who spent more time on the websites or visit it often.
- leads who are the working professionals.
- leads which originate from "Add form".

The company should not make calls to the leads who match the below criteria as these are not likely to get converted:

- lead origin is "Landing Page Submission"
- leads whose Specialization was "Not Specified"
- leads whose Occupation is "Unknown"
- leads who chose the option of "Do not Email" as "Yes"
- leads who opt for a free copy of "Mastering the Interview" book.

**Business requirement and Recommendations**

- Lead score value for each lead was calculated. Using this score, the “Hot Leads” for X Education were identified.
- Target was to obtain lead conversion rate to be around 80%. Using the final model, the specificity in the test dataset was 77.2% and accuracy is 79.18%.
- By reducing the cut off limit, the business target can be achieved. Depending on the business requirement, we can increase or decrease the probability threshold value which will in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.
- Increasing the cut off limit will ensure that the business focuses on those customers who have very high probability of conversion or “Hot Leads”.
- As observed from the driver variables, the website plays an important role in determining the conversion rate. Hence, it must be more engaging, and the content must be up to date such that the leads spend more time and visits it often.