X Education - Lead Scoring Case Study

Detection of Hot Leads to concentrate more of marketing efforts on them, improving conversion rates for X Education

Team Members: Pratik Kumar Routray

Table of Contents

Background of X Education Company

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.

Problem Statement & Objective of the Study

Problem Statement:

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we 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 has given a ballpark of the target lead conversion rate to be around 80%.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are grouped based on their propensity or likelihood to convert.
- This results in a focused group of hot leads.

Better Communication

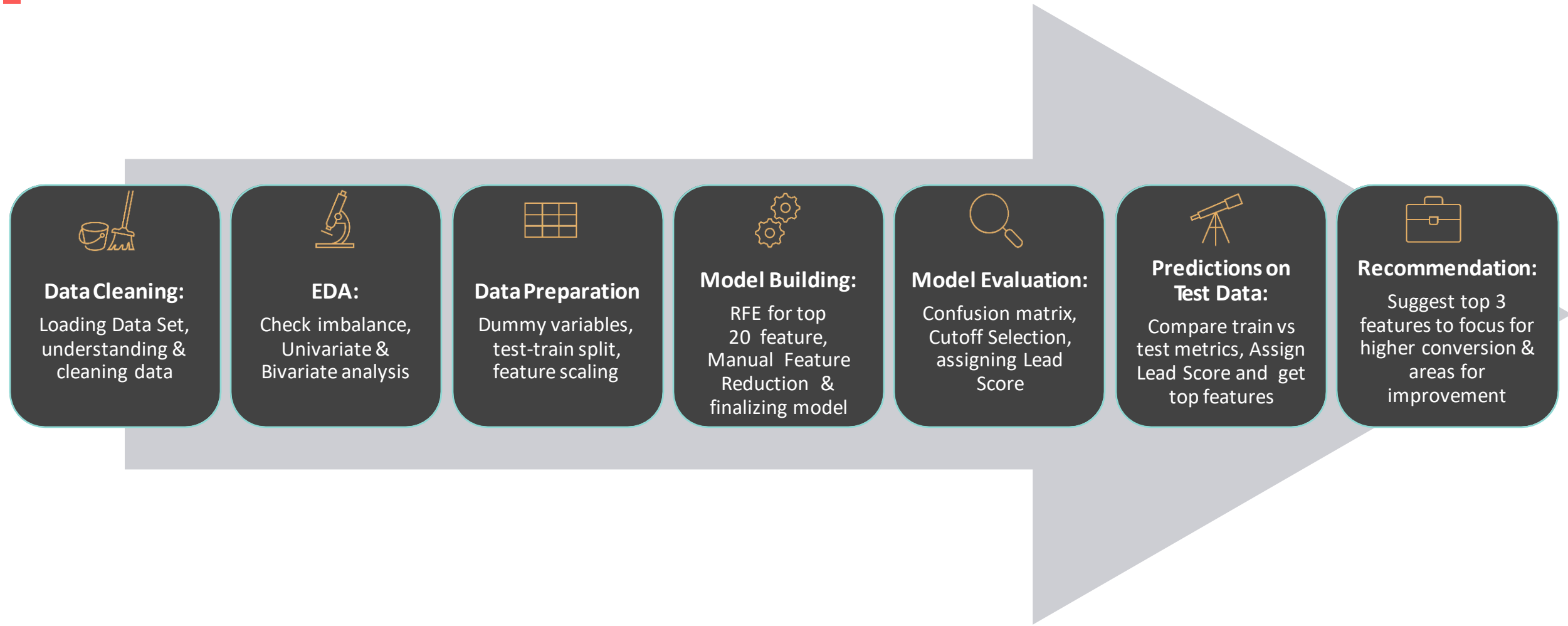
- We could have a smaller pool of leads to communicate with, which would allow us to have a greater impact.

Boost Conversion

- We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to convert.

Since we have a target of 80% conversion rate, we would want to obtain a high **sensitivity** in obtaining hot leads.

Analysis Approach



Data Cleaning

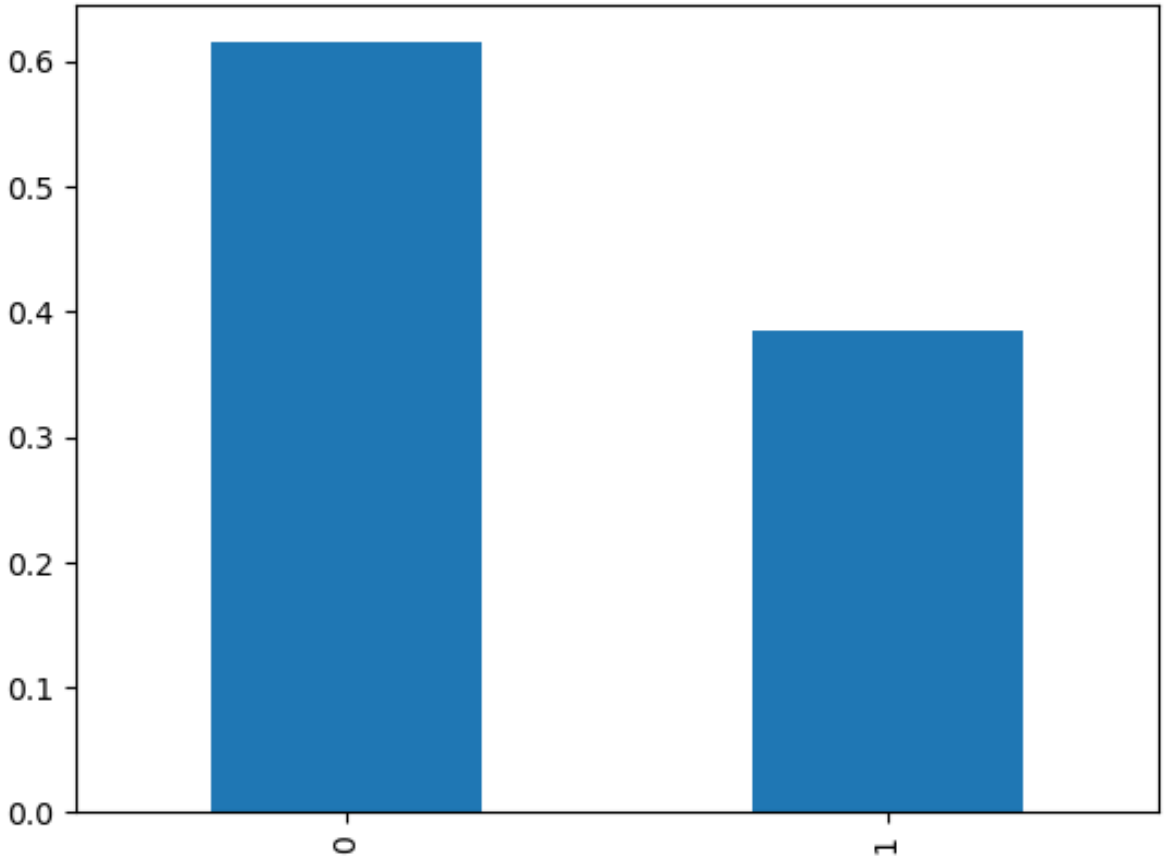
- **"Select"** level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

Data Cleaning

- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in **TotalVisits** and **Page Views Per Visit** were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to “Others”.
- Binary categorical variables were mapped.
- Other cleaning activities were performed to ensure data quality and accuracy.
    - Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google)

EDA

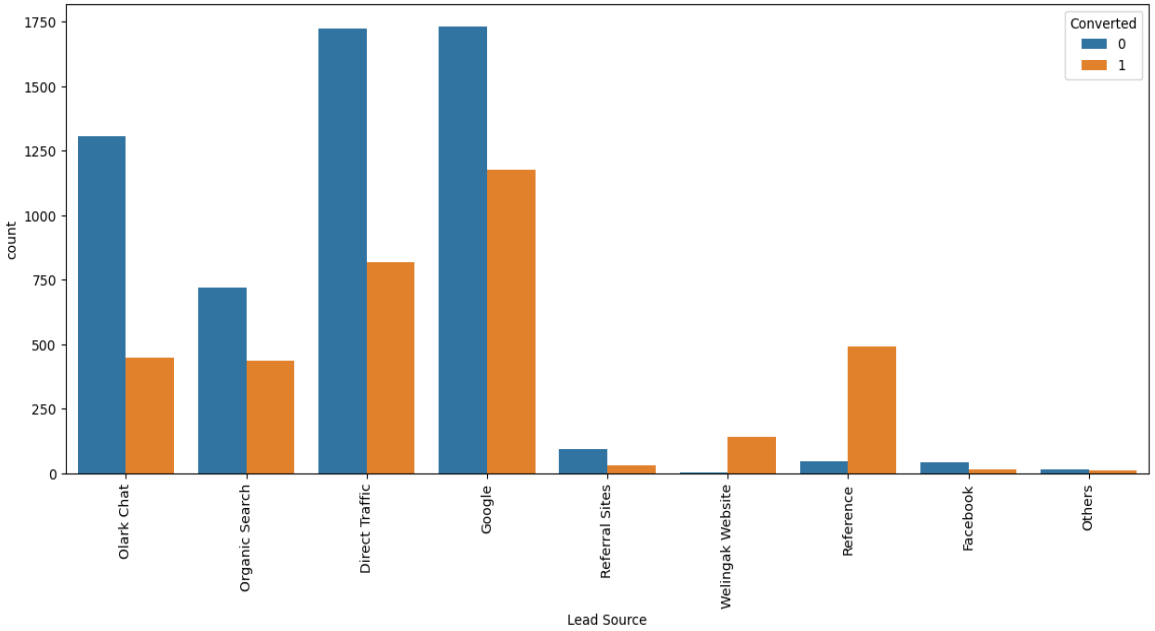
- Data is imbalanced while analyzing target variable.



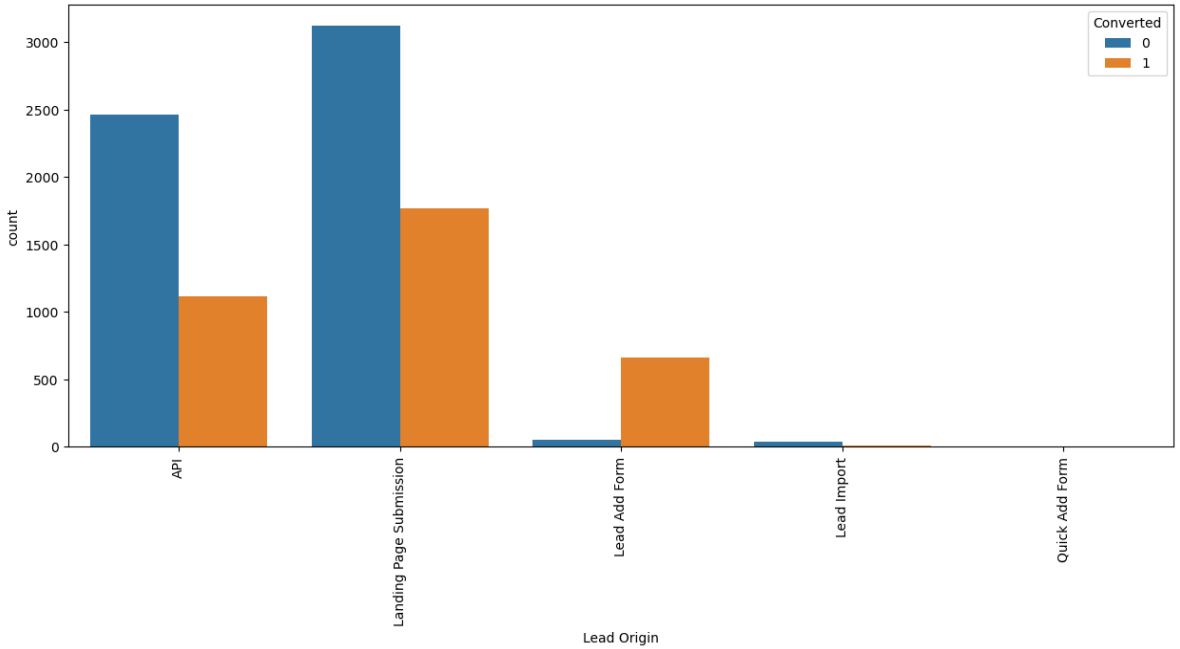
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

EDA

- Univariate Analysis - Categorical Variables



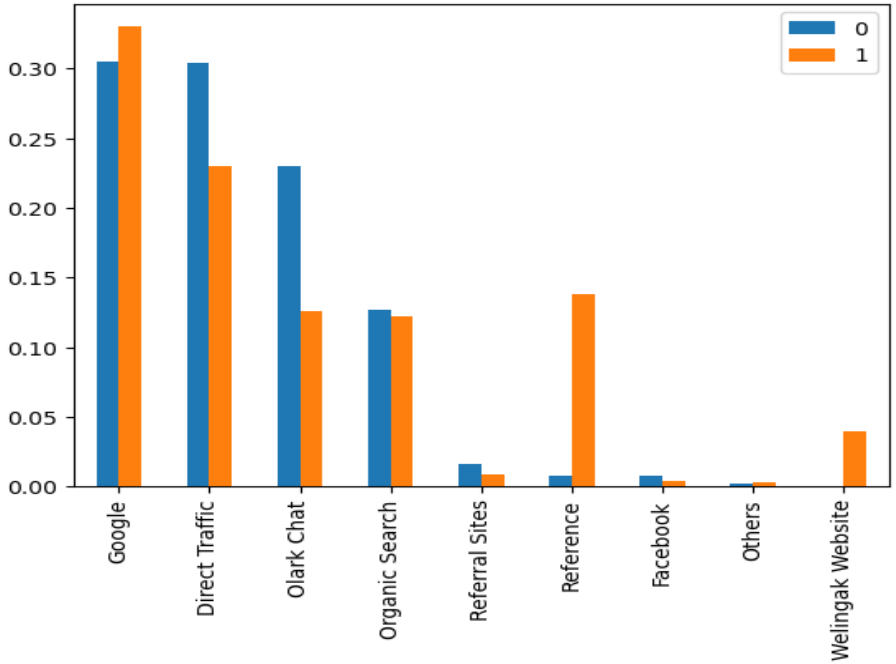
Most of the leads are from google but reference has the highest conversion rate

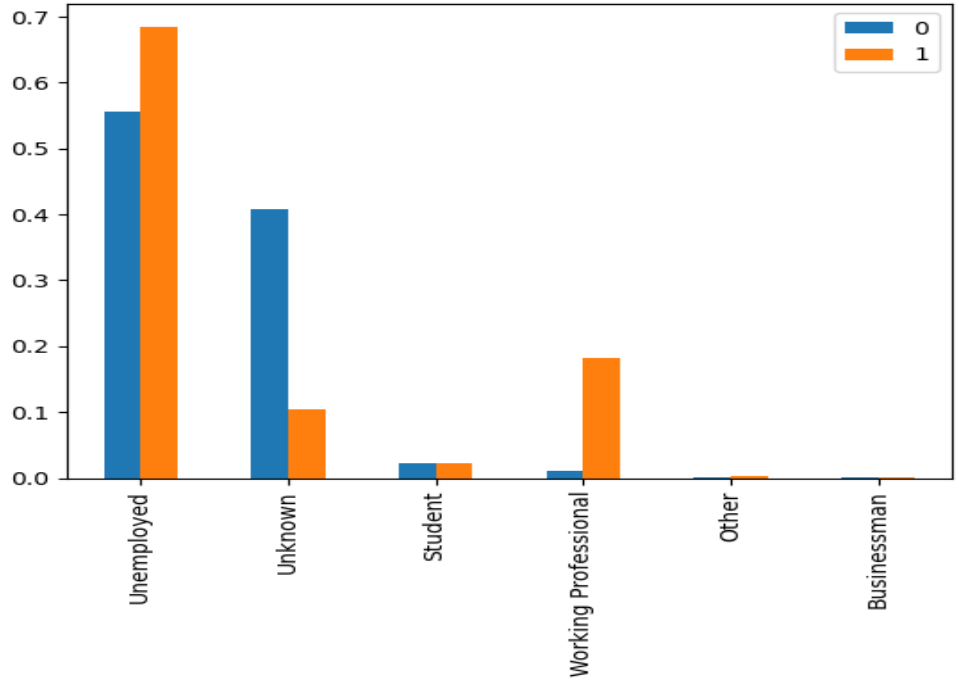


Similarly lead origin highest from landing page submission

EDA

- **Bivariate Analysis – Categorical Variables**

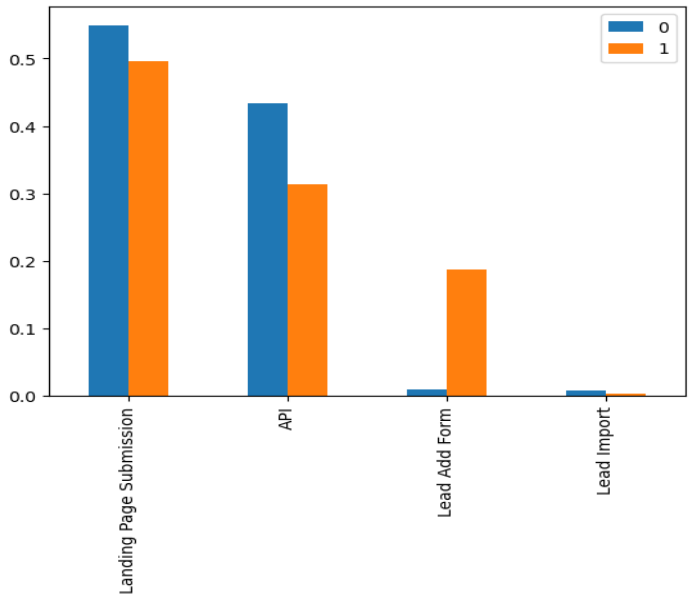


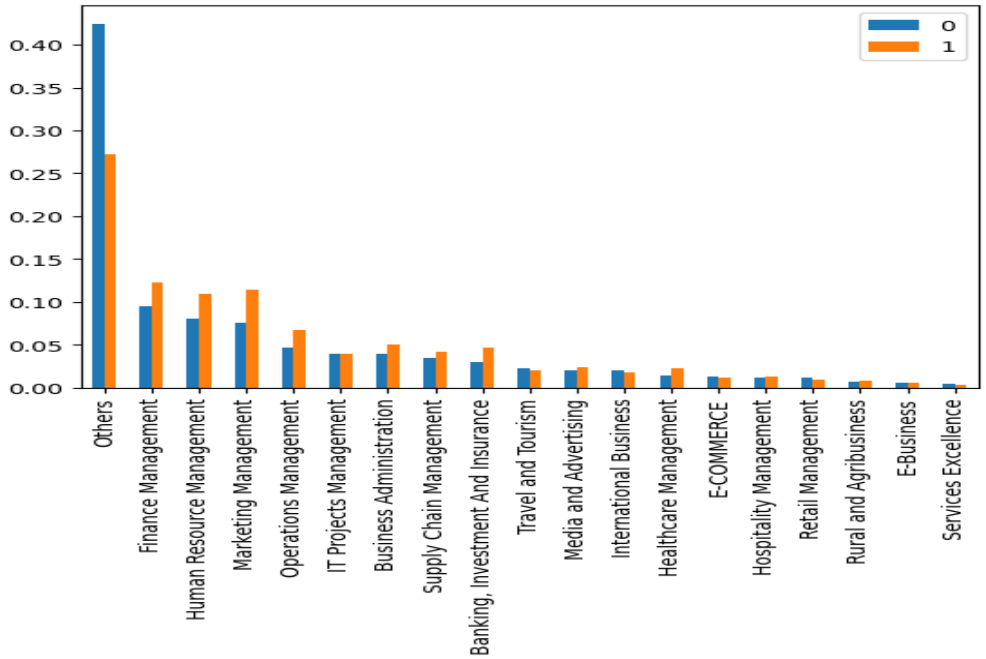


- **Lead Origin: Similarly, here we can see that leads generated through google and reference has higher conversion rate than others in terms of percentage.**
- **Current_occupation: customers of unemployed and working professionals are highly converted in terms of percentage.**

EDA

- **Bivariate Analysis – Categorical Variables**

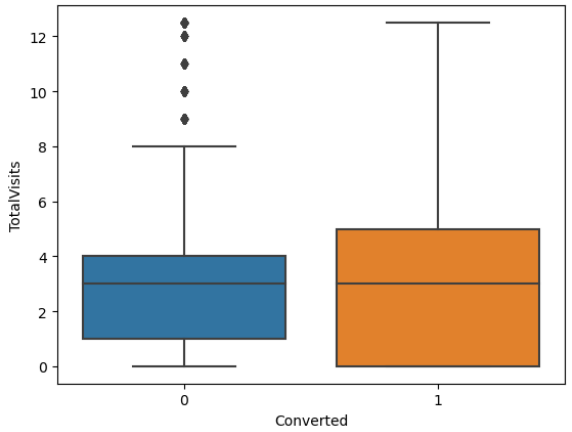


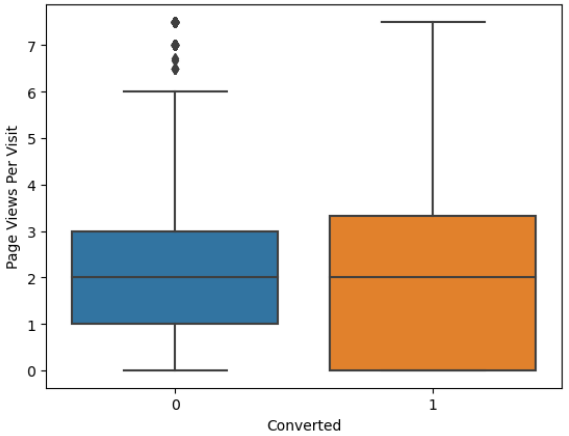


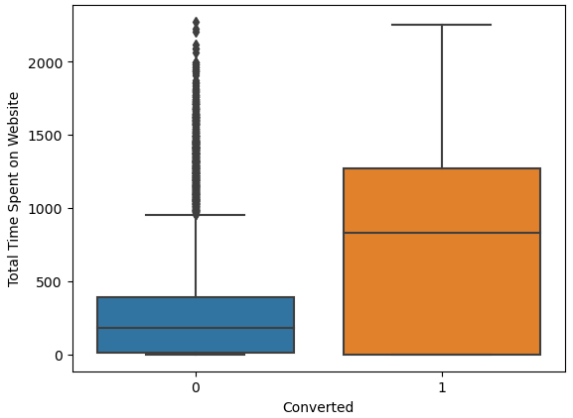
- **Highest customers are generated through landing page submission but from lead add form the customers are higly likely to be converted.**

- **Courses of Financial management marketing management or other management courses are mostly chosen by customers or then are highly likely to be converted,**

EDA - Bivariate Analysis for Numerical Variables

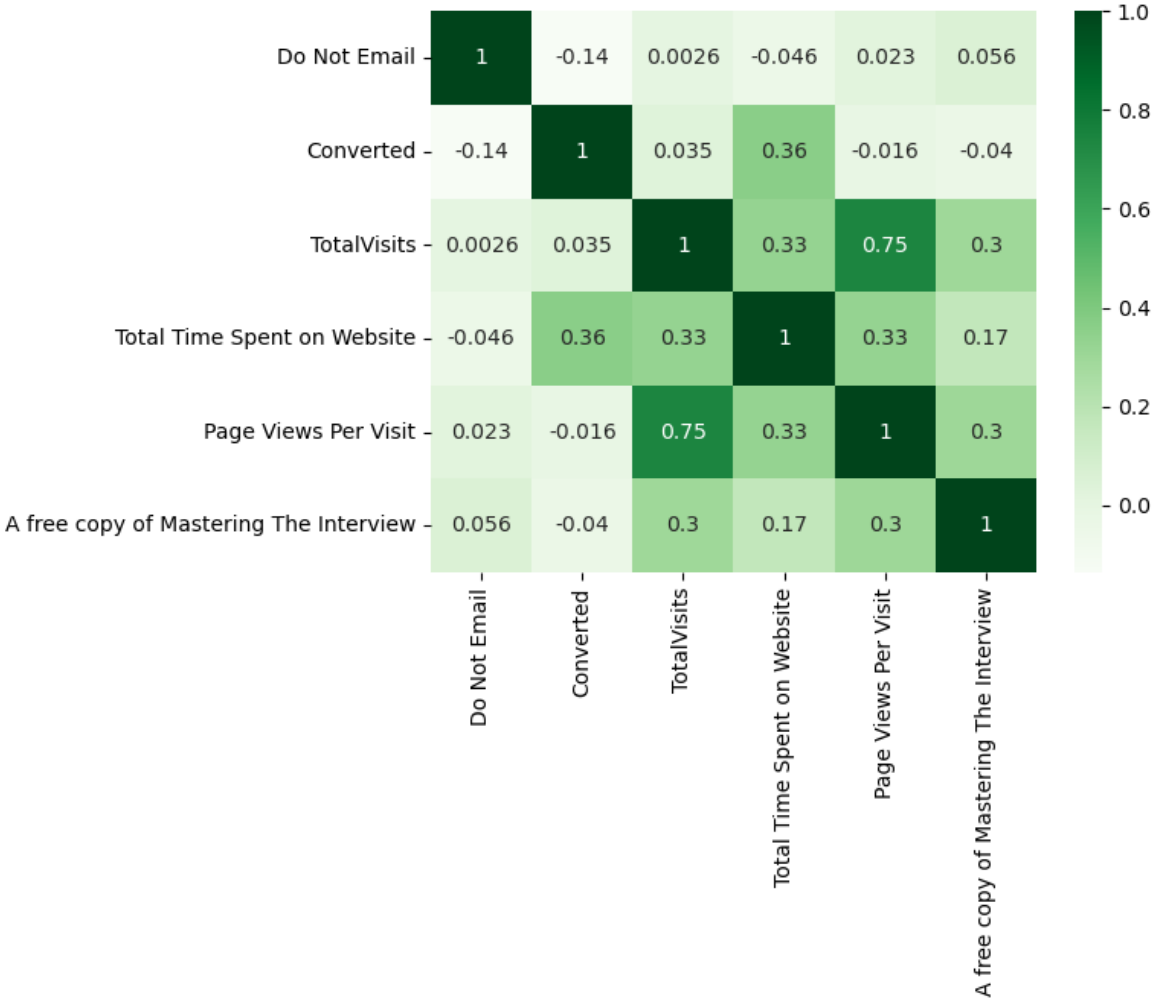






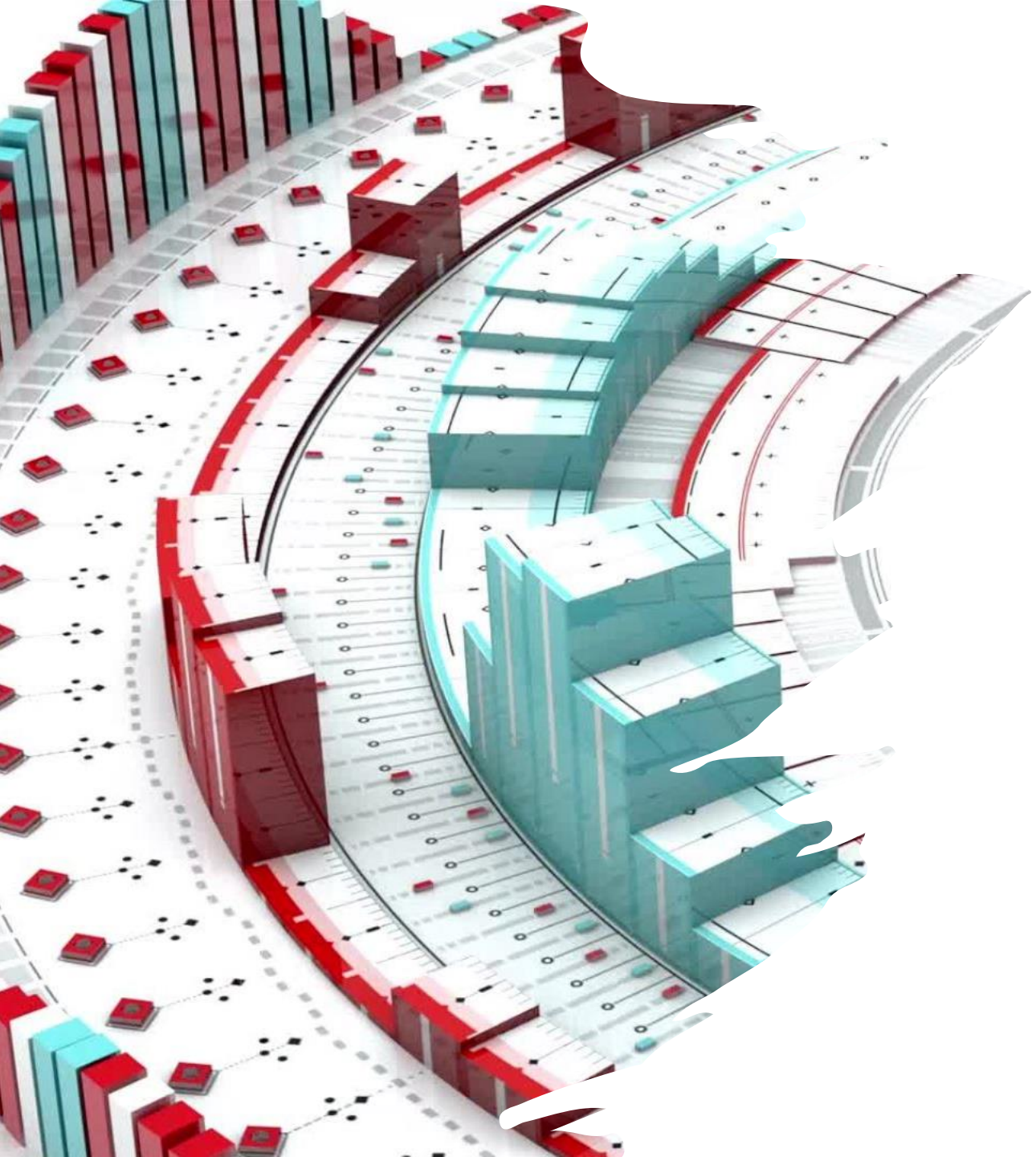
- Past Leads who **spends more time on the Website** have a higher chance of getting successfully converted than those who spends less time as seen in the **box-plot**

EDA - Heat map for the Numerical variables to check the correlation among the Variables.



Data Preparation before Model building

- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation
- Splitting Train & Test Sets
  - 80:20 % ratio was chosen for the split
- Feature scaling
  - MinMaxScaler method was used to scale the features.



Model Building

- **Feature Selection**
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform **Recursive Feature Elimination** (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome
  - Pre RFE – 51 columns & Post RFE – 20 columns

Model Building

- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 4 looks stable after four iteration with:
  - significant p-values within the threshold (p-values < 0.05) and
  - No sign of multicollinearity with VIFs less than 5
- Hence, **logmf** will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

Model Evaluation

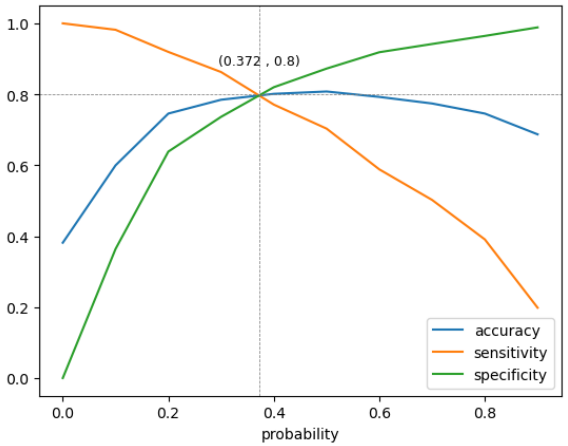
Train Data Set

It was decided to go ahead with 0.372 as cutoff after checking evaluation metrics coming from both plots

Confusion Matrix & Evaluation Metrics with 0.372 as cutoff

```
Confusion Matrix
[[3674  898]
 [ 587 2233]]
```

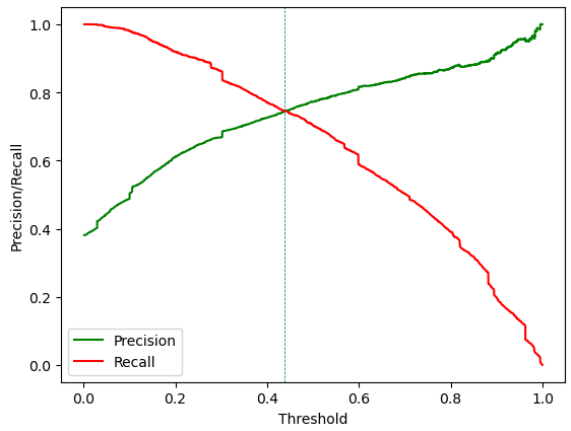
```
True Negative                  :  3674
True Positive                  :  2233
False Negative                 :  587
False Positve                  :  898
Model Accuracy                 :  0.799
Model Sensitivity              :  0.792
Model Specificity              :  0.804
Model Precision                :  0.713
Model Recall                   :  0.792
Model True Positive Rate (TPR) :  0.792
Model False Positive Rate (FPR) :  0.196
Model False Negative Rate (FNR) :  0.208
```



Confusion Matrix & Evaluation Metrics with 0.438 as cutoff

```
Confusion Matrix
[[3851  721]
 [ 714 2106]]
```

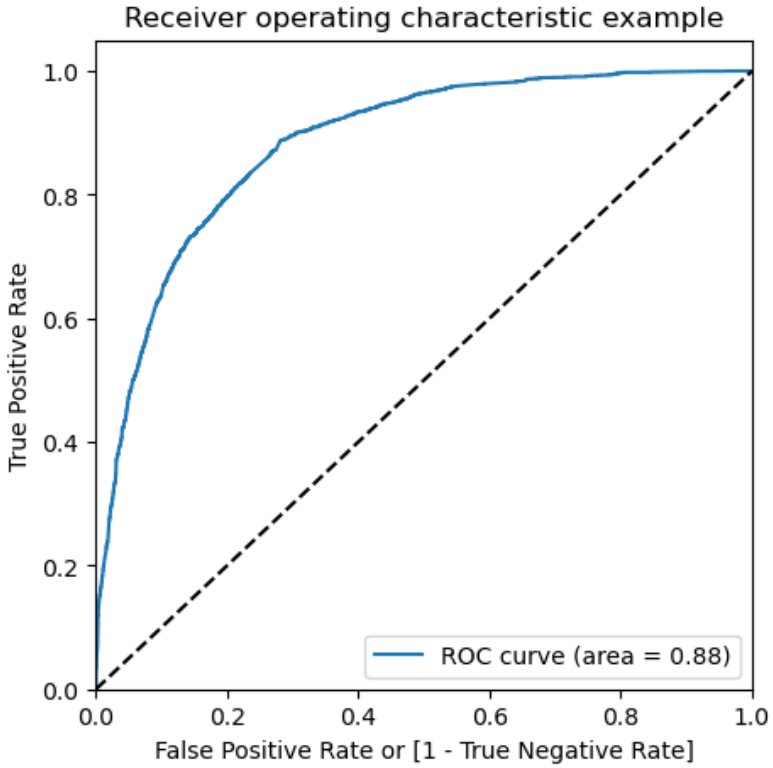
```
True Negative                  :  3851
True Positive                  :  2106
False Negative                 :  714
False Positve                  :  721
Model Accuracy                 :  0.806
Model Sensitivity              :  0.747
Model Specificity              :  0.842
Model Precision                :  0.745
Model Recall                   :  0.747
Model True Positive Rate (TPR) :  0.747
Model False Positive Rate (FPR) :  0.158
Model False Negative Rate (FNR) :  0.253
```



Model Evaluation

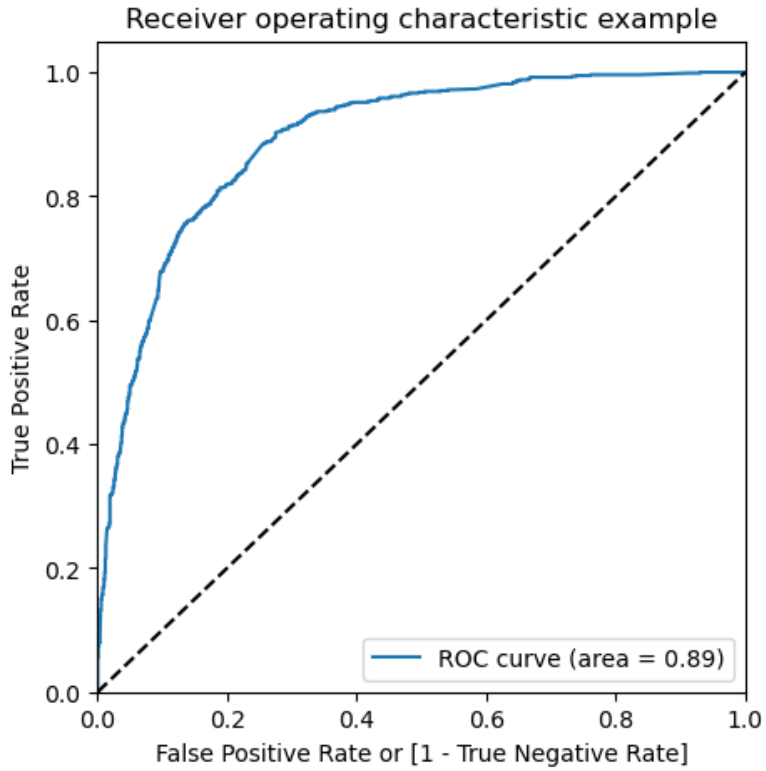
ROC Curve – Train Data Set

- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve – Test Data Set

- Area under ROC curve is 0.89 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation

Confusion Matrix & Metrics

Train Data Set

```
Confusion Matrix
[[3674  898]
 [ 587 2233]]

_____________________________________________

True Negative                   :  3674
True Positive                   :  2233
False Negative                  :  587
False Positve                   :  898
Model Accuracy                  :  0.799
Model Sensitivity               :  0.792
Model Specificity               :  0.804
Model Precision                 :  0.713
Model Recall                    :  0.792
Model True Positive Rate (TPR)  :  0.792
Model False Positive Rate (FPR) :  0.196
Model False Negative Rate (FNR) :  0.208
```

Test Data Set

```
_____________________________________________

Confusion Matrix
[[887 220]
 [134 607]]

_____________________________________________

True Negative                   :  887
True Positive                   :  607
False Negative                  :  134
False Positve                   :  220
Model Accuracy                  :  0.808
Model Sensitivity               :  0.819
Model Specificity               :  0.801
Model Precision                 :  0.734
Model Recall                    :  0.819
Model True Positive Rate (TPR)  :  0.819
Model False Positive Rate (FPR) :  0.199
Model False Negative Rate (FNR) :  0.181
```

- Using a cut-off value of 0.372, the model achieved a **sensitivity** of 79.9% = **80 % in the train set** and **80.80% in the test set**.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target **sensitivity of around 80%**.
- The model also achieved an **accuracy of around 80%**, which is in line with the study's objectives.

Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
  - Total Time Spent on Website 4.478814
  - Lead Source_Welingak Website 3.082976
  - Lead Origin_Lead Add Form 3.045473
  - Last Activity_Others 1.854357
  - Last Activity_SMS Sent 1.834202
  - Lead Source_Olark Chat 1.311766
  - TotalVisits 0.717152
  - Last Activity_Unreachable 0.646639
  - Last Activity_Email Opened 0.597641
- We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:
  - Specialization_Others -1.112613
  - What is your current occupation_Unknown -1.356362
  - Do Not Email -1.364515

Recommendation based on Final Model

- **To increase our Lead Conversion Rates**
  - For tailored marketing efforts, pay attention to features with high correlation coefficients.
  - Create tactics to entice top-performing lead sources to send you high-quality leads.
  - Adapt communication channels based on the impact of lead engagement.
  - Communicate with working professionals in a relevant way.
  - On the Welingak website, further spending can be made on things like advertising.
  - Rewards/discounts for supplying references that result in leads promote supplying more references.
  - Working professionals should be aggressively targeted because they convert well and are more likely to have the money to pay higher fees.
- **To identify areas of improvement**
  - Review the specialized options' negative coefficients.
  - Look for areas for improvement in the landing page submission procedure.

Thank You!