Lead Scoring Case Study Analysis

Business Problem

- An education company "***X Education***" is in the business of selling online courses to industry professionals.
- Company markets their courses on websites and search engines. Once people land on the website and fill the form mentioning their personal information such as email address or phone number they become "***lead***".
- These "***leads***" are then contacted by the company's sales representative through calls or e-mails. Typically, out of the total acquired leads, only **30%** are converted to ***paying customers***.

Business Objective

- Since the lead conversion rate is low **(30%)** the company wants to save resources and improve the lead conversion efficiency.
- For this the company needs help with identifying the highly promising prospects also termed as "***Hot Leads***" that are very highly probable to become a client . This in turn ensures the conversion rate goes high and lead conversion process turns more efficient.
- For this "**X Education"** expects to build a logistic regression model that would assign score to each of the leads between ***0-100*** *which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.*
- The **CEO** has given the ballpark target of improving the lead conversion rate to **80%** from the current **30%.**

Solution Steps

- Data exploration
- Data preprocessing
  - Data cleaning: Removing redundant values and columns
  - Outliers and missing values treatment
- Exploratory Data Analysis and Visualization
- Model building
  - Feature Scaling and transformation
  - Feature Selection using RFE, VIF and p-values
  - Model Training and Evaluation
- Business Result verification
- Recommendations

Data Exploration and Analysis

Data observation

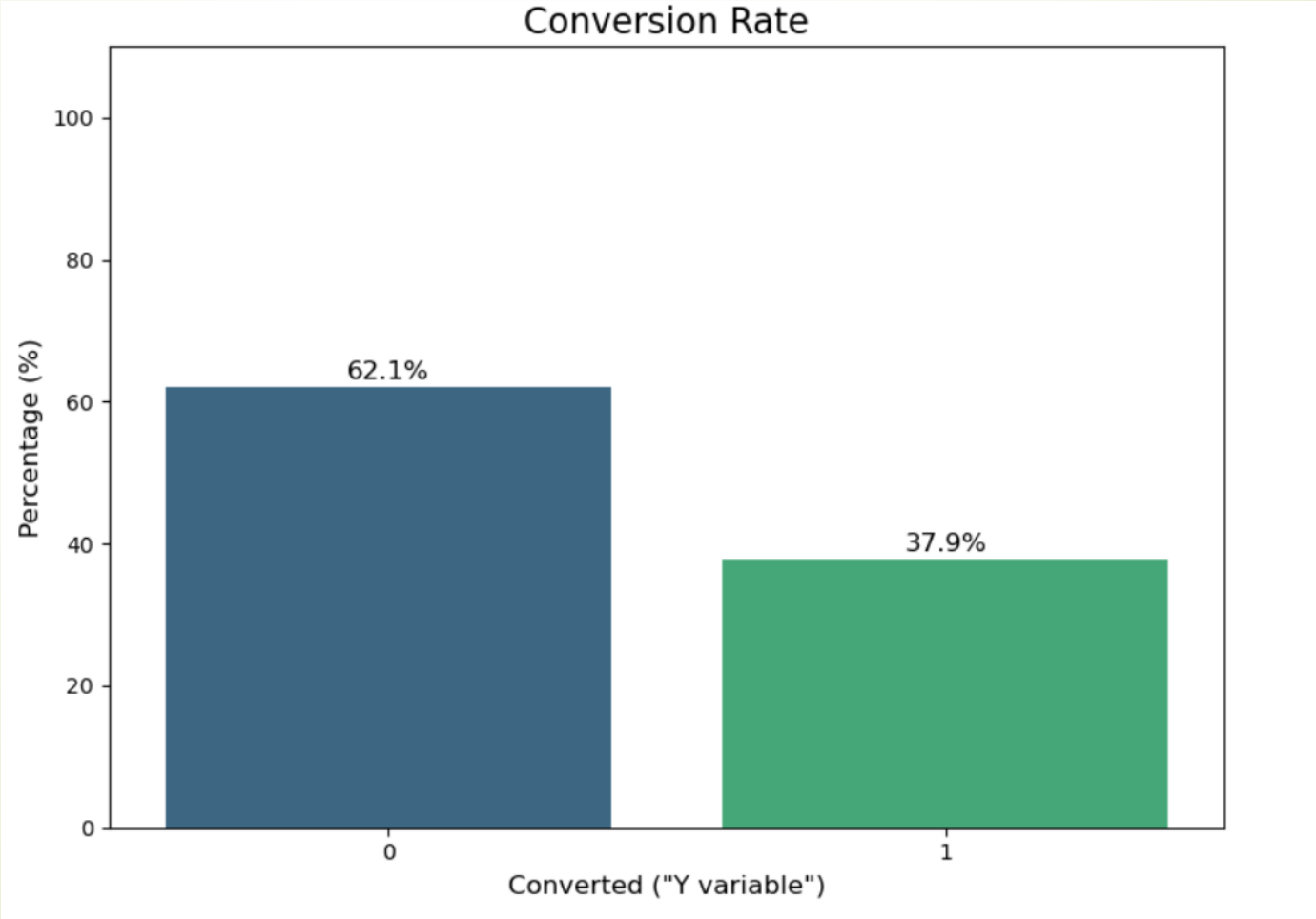
- A total of 9,240 records were available with about 37 different parameters mentioning available to assess.

```
[5]: #Viewing the shape of the dataset
     lead.shape

[5]: (9240, 37)
```

- A split of different types of data was available for analysis consisting of numeric and object types in general.

Split of Target Variable

- The target variable here is the **"Converted"** column. The provided dataset depicts the typical conversion rate of **30%** with a split of **37.9%** and **62.1%** between the converted and non-converted leads respectively.



Irregularities in data

- "**Select**" is a value that indicates non-selection of any options which is present in few of the columns.

```
Columns that have value 'Select':
        Specialization:1942
        How did you hear about X Education:5043
        Lead Profile:4146
        City:2249
```

- Certain columns are uni-valued

```
Magazine:['No']
Receive More Updates About Our Courses:['No']
Update me on Supply Chain Content:['No']
Get updates on DM Content:['No']
I agree to pay the amount through cheque:['No']
```

Data preprocessing

- The value “**Select”** is replaced as “**NaN**” value.
- Uni-valued columns are dropped.
- **Prospect ID** and **Lead Number** are dropped since they are mere undexes
- The Country column has about 38 countries. To reduce the difficulties in the analysis the values are segregated in three buckets namely **India**, **Not Given**, **Other** Country.

Missing values and Treatment

- Percentage of **missing values** observed for **columns** ranged from **0.39%** for **Lead Source** to **78.46%** for **How did you hear about X Education**.
- Percentage of **missing values** observed for each **row** is less than **1%.**

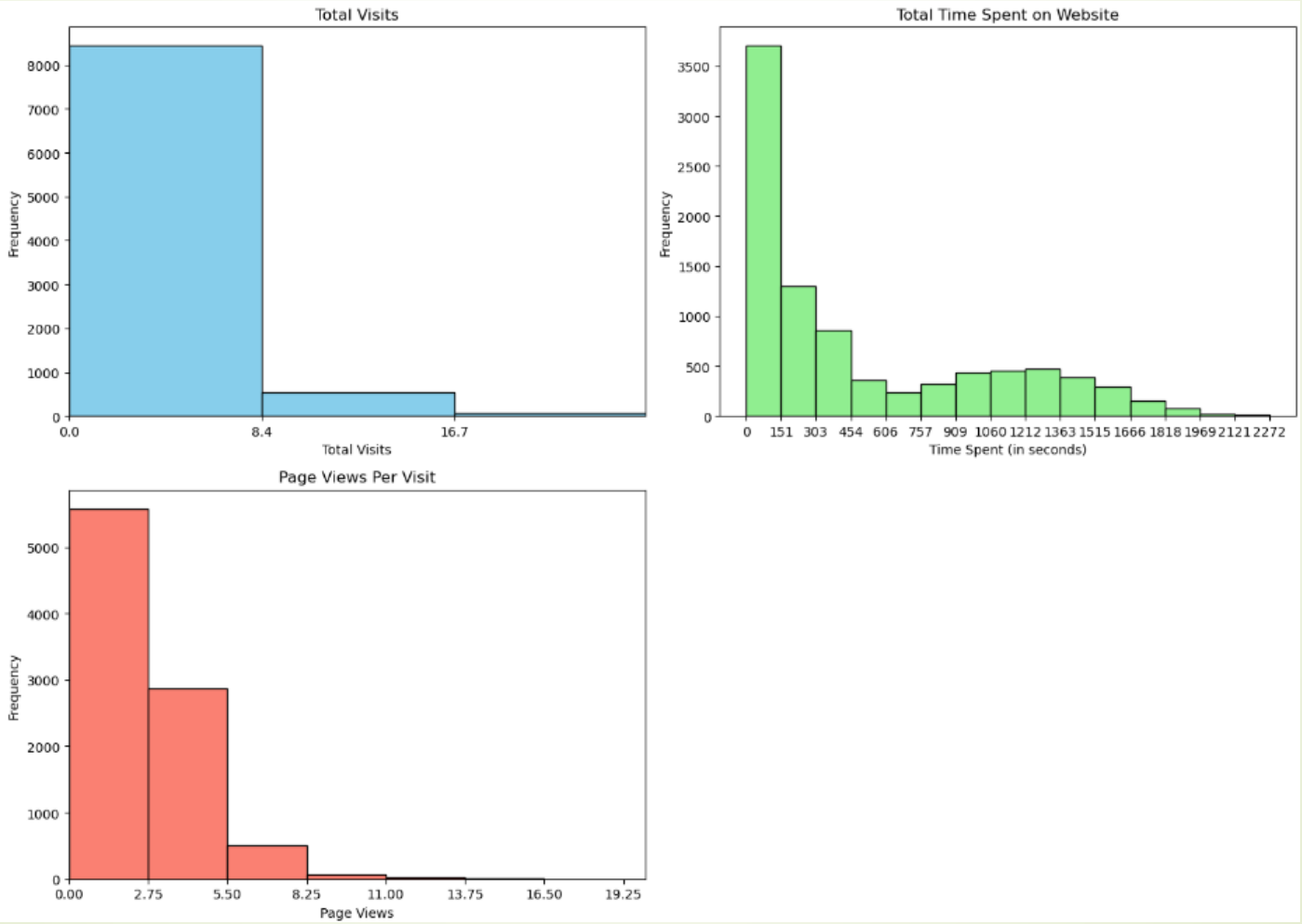
```
Lead Source : 0.39 %
TotalVisits : 1.48 %
Page Views Per Visit : 1.48 %
Last Activity : 1.11 %
Country : 26.63 %
Specialization : 36.58 %
How did you hear about X Education : 78.46 %
What is your current occupation : 29.11 %
What matters most to you in choosing a course : 29.32 %
Tags : 36.29 %
Lead Quality : 51.59 %
Lead Profile : 74.19 %
City : 39.71 %
Asymmetrique Activity Index : 45.65 %
Asymmetrique Profile Index : 45.65 %
Asymmetrique Activity Score : 45.65 %
Asymmetrique Profile Score : 45.65 %
```

Missing values and Treatment

- Following the thumb rule, columns with more than **40%** missing values were dropped since they would hinder further analysis also those values representing missing values were replaced with "**Not Given**" if they were of insignificant count.
- Dropping rows with missing values would result in a loss of **5.889%** of data and hence were also dropped.
- Additionally, certain redundant columns were also dropped.

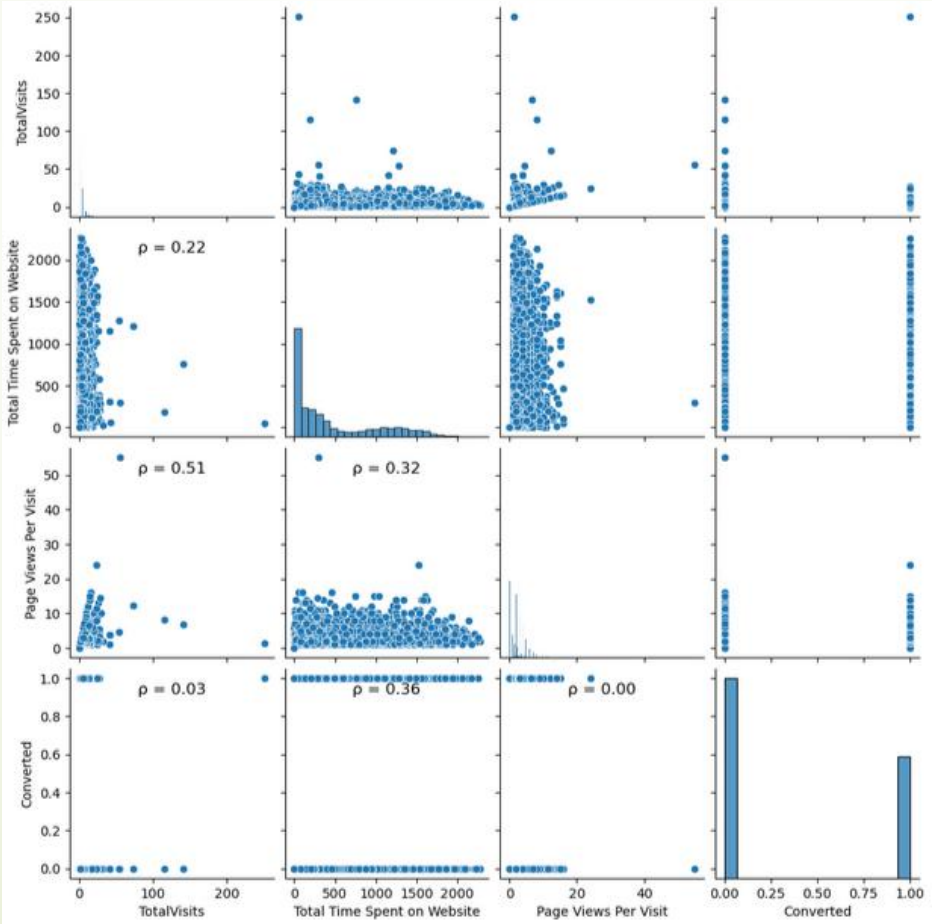
Exploratory Data Analysis

- While the courses are marketed online, a customer makes an approximate **Total of 0 – 8.4 visits** spending about **1-454 seconds** on website with an **majority viewing 5.5 pages**.



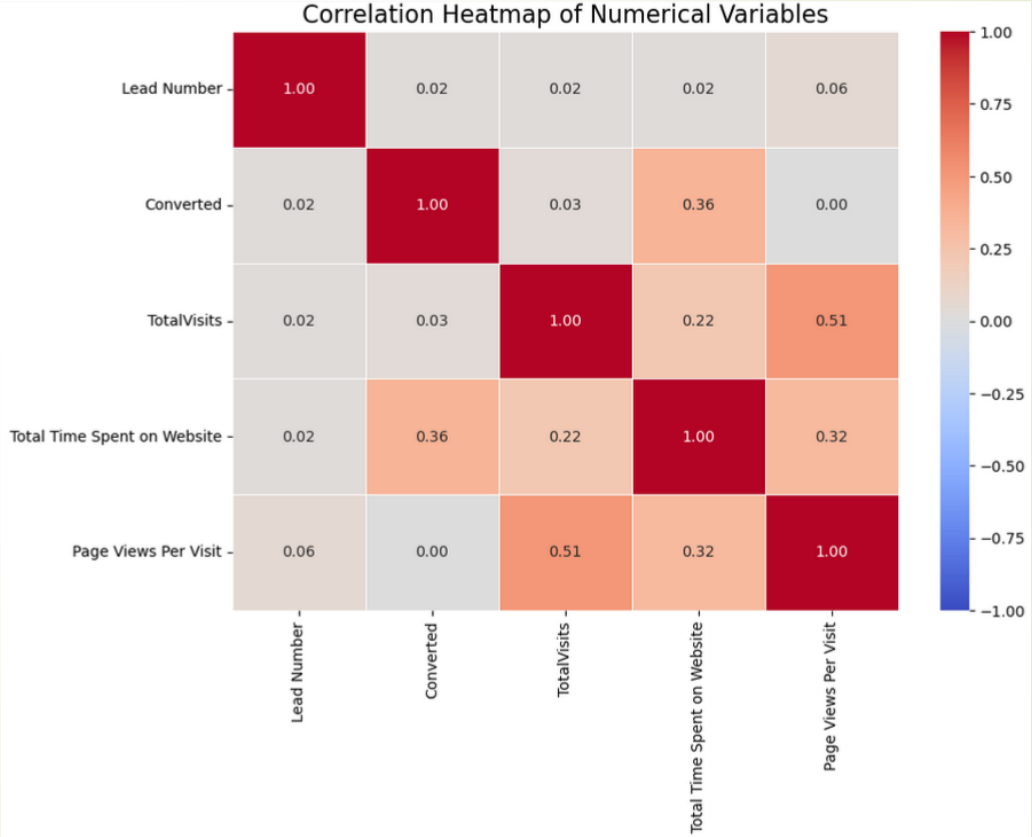
Exploratory Data Analysis

- While plotting the pairplots and the correlation between Total Visits, Total time spent on website and page views per visit and Converted there is an observed **positive correlation** of **0.51** between **Total Visits** and **Page Views per visit** and a **positive correlation** of **0.36** between **Total time spent on website** and **Converted**.



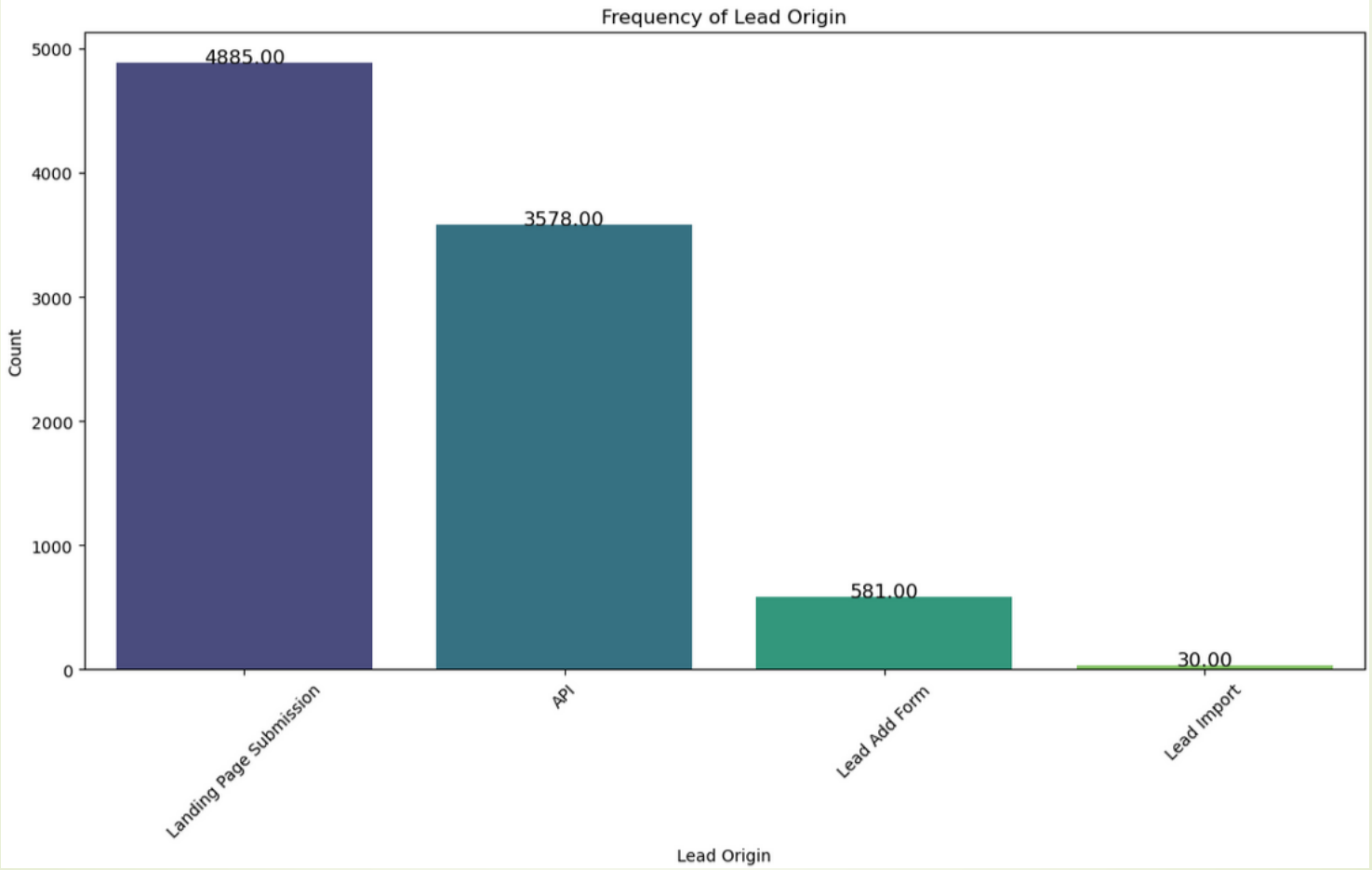
Exploratory Data Analysis

- The correlation plot conveys two major points: -
  - **Total Time Spent on Website** has a **positive correlation** with correlation coefficient of **0.36** with the **Converted** variable.
  - **Total Visits** and **Page Views per Visit** have **positive correlation** with correlation coefficient of **0.51** hinting at multicollinearity.



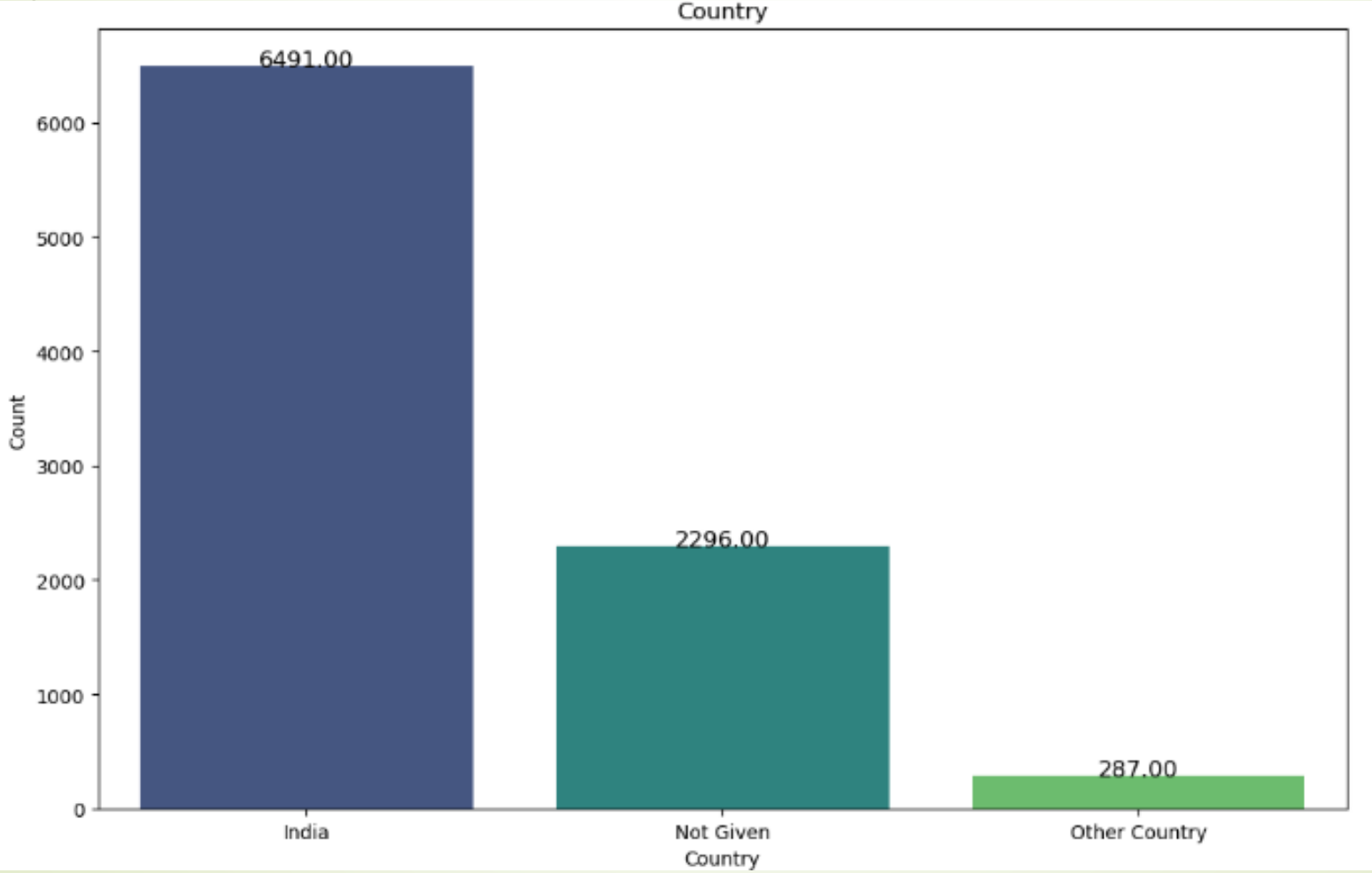
Exploratory Data Analysis

- To identify **the Landing Page Submission** identified about **4885** customer as Lead and **API** identified about **3578** leads.



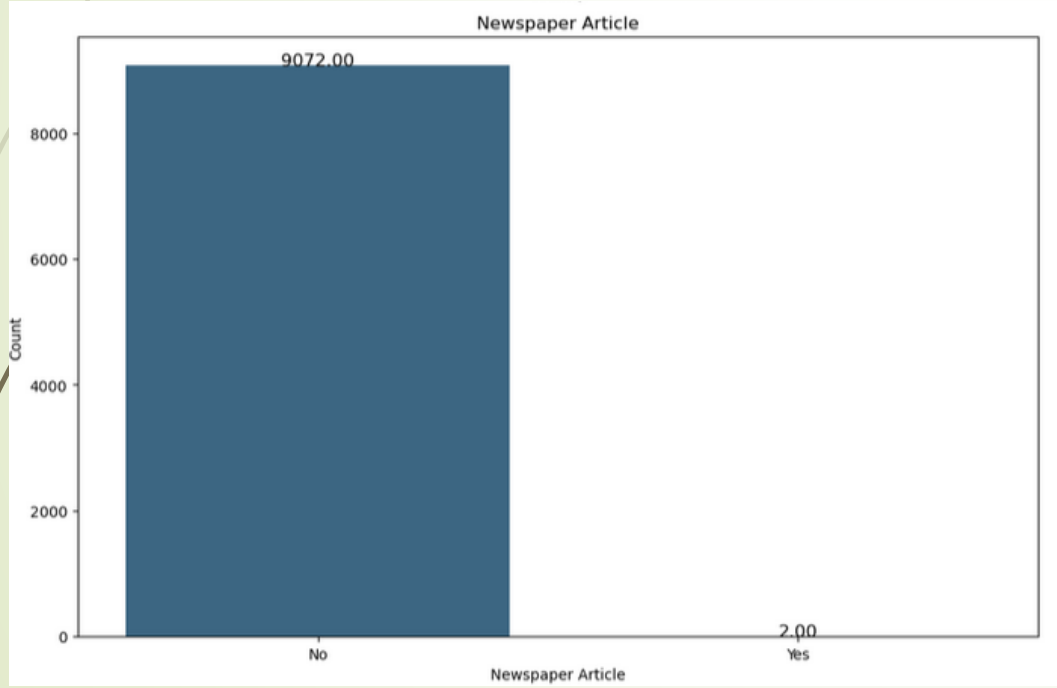
Exploratory Data Analysis

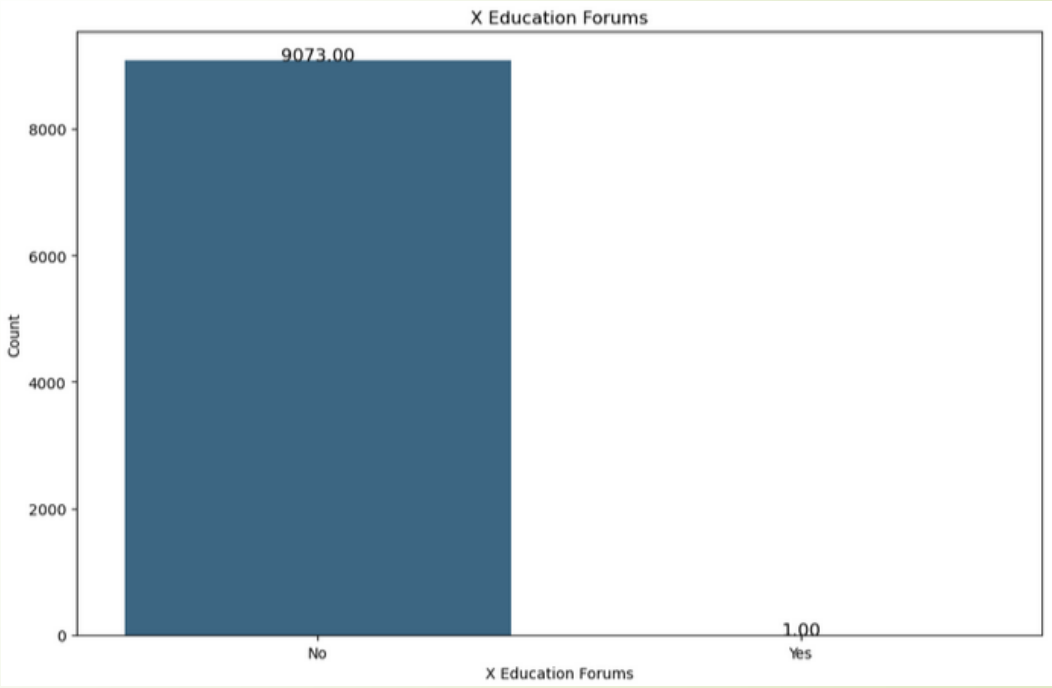
- **India** tops the customer list with a total of **6491** customers whereas only **287** are from **other** countries while the remaining **2296** customers have not provided their country information.



Exploratory Data Analysis

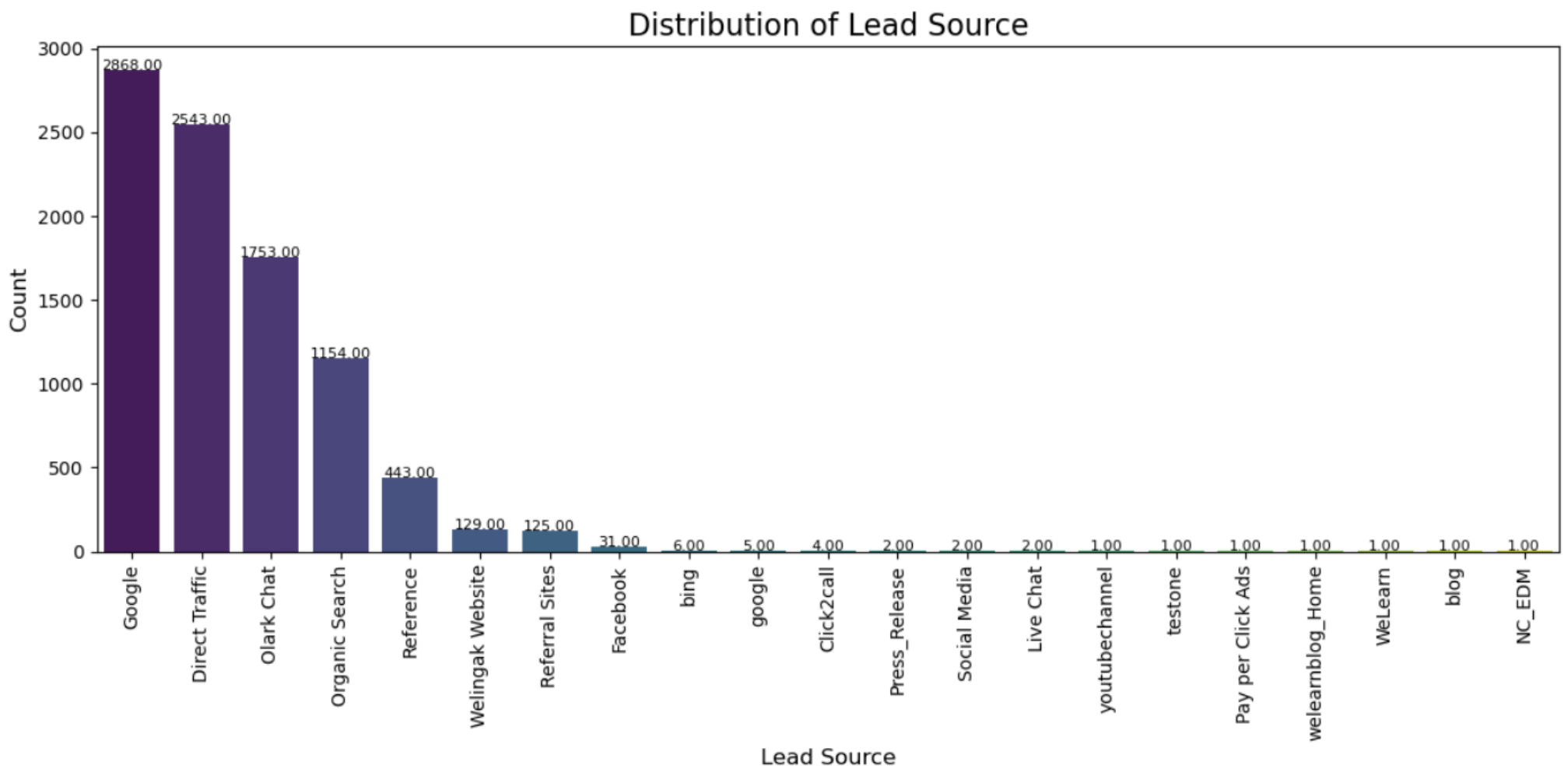
- Advertisement to raise awareness about the course were done at multiple channels or mediums such as **Newspaper Article**, **X Education Forum**, **Digital Advertisement etc.,** but a very little chunk of customers had seen those advertisement. For example, only **2** customers had seen the **Newspaper Article** and only **1** had seen advertisement on **X Education Forums**.





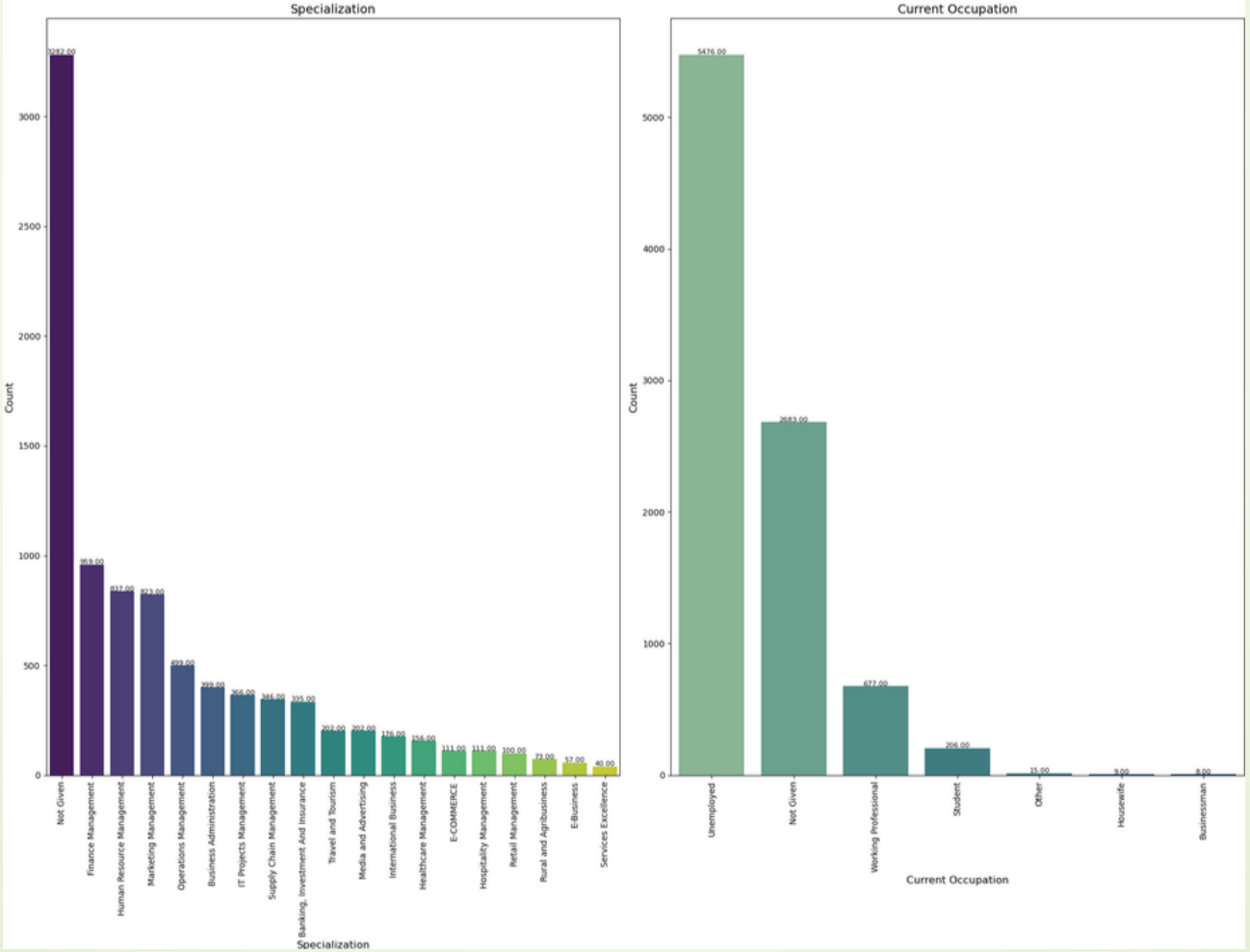
Exploratory Data Analysis

- Out of all the different **Lead Sources**, **Google** was the highest with about **2868** lead source followed by **Direct Traffic** i.e. website visit with about **2543** sources.



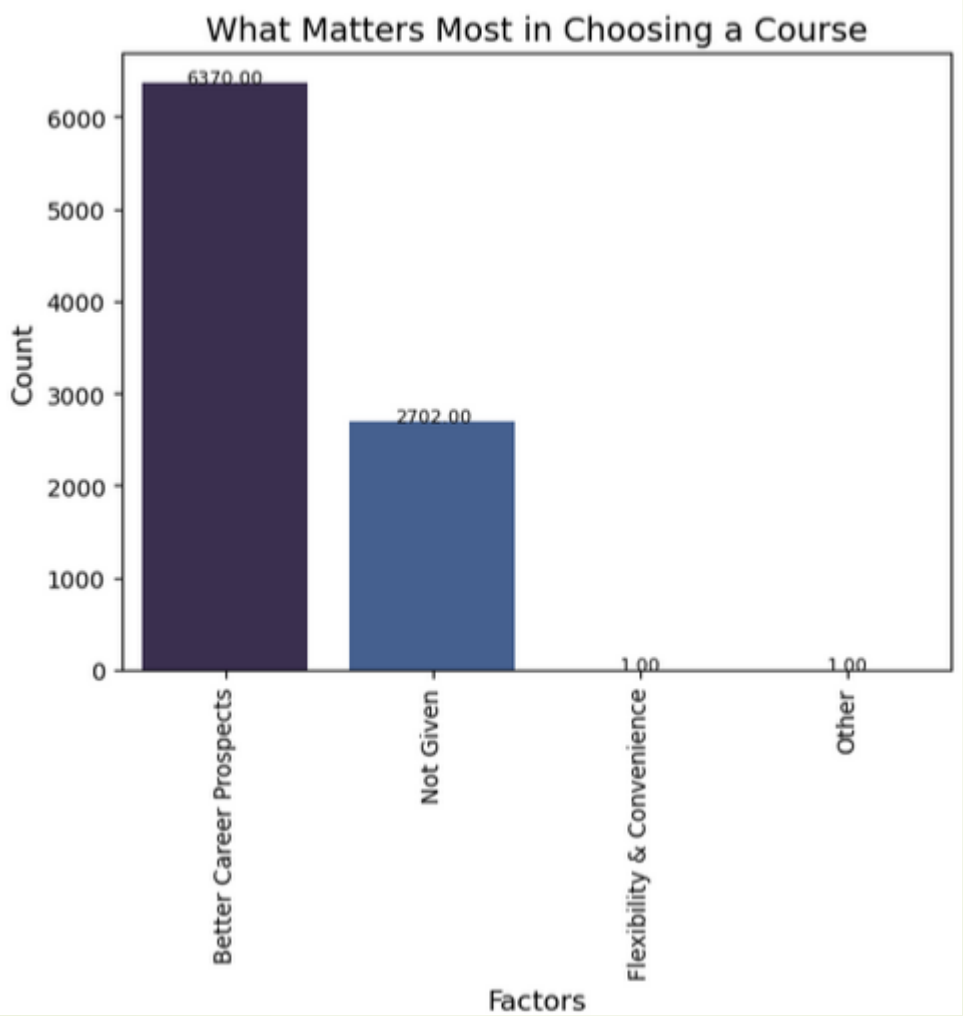
Exploratory Data Analysis

- Considering the specialization background, at the top are **959 Finance Management, 837** from **Human Resource Management** indicating people in Management specialization check out X Educations course while the major chunk of **5476** customers is currently **unemployed.**



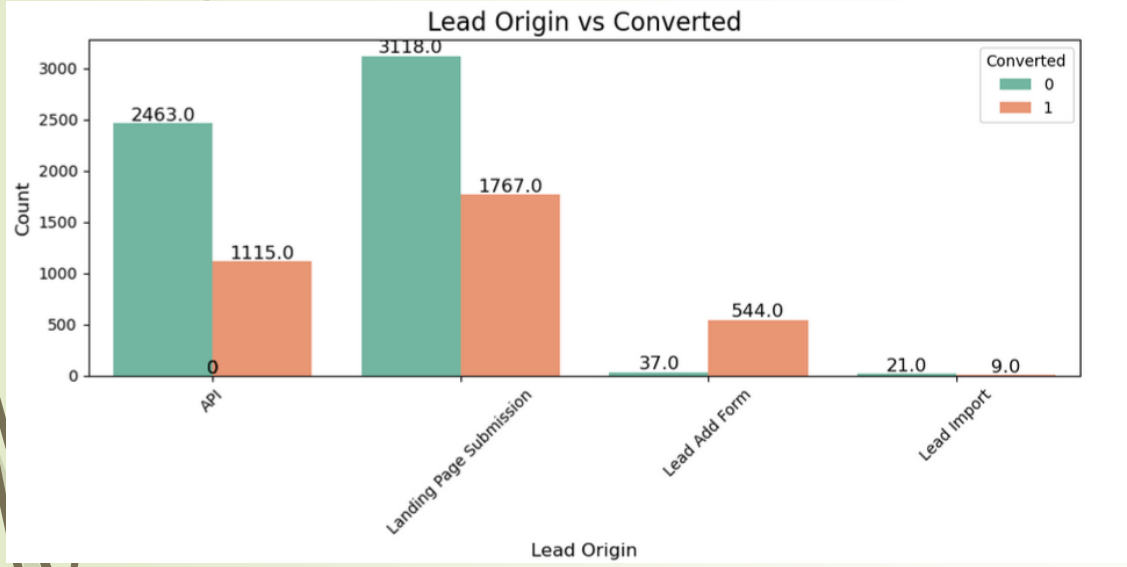
Exploratory Data Analysis

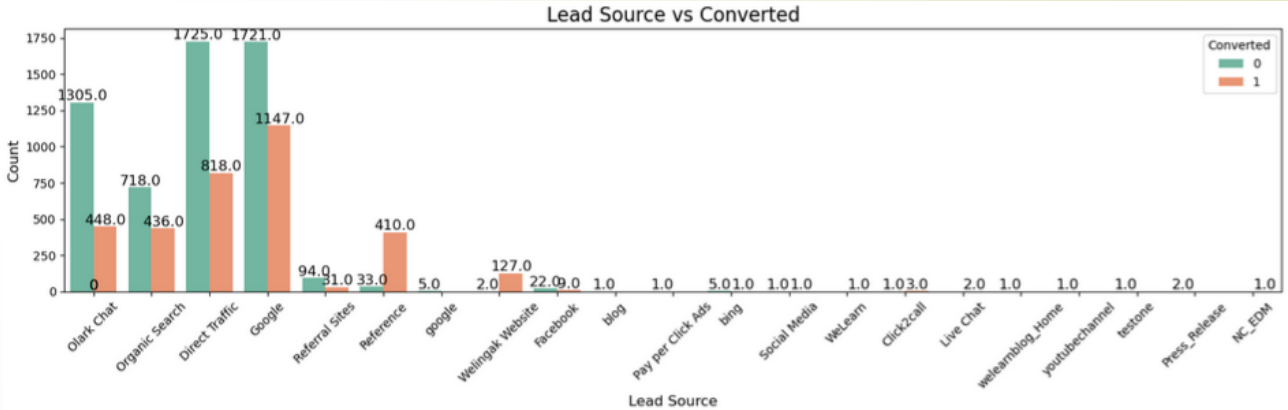
- Also, of all the customers **6370** stated the most important thing for selecting a course is **Better Career Prospect.**



Exploratory Data Analysis

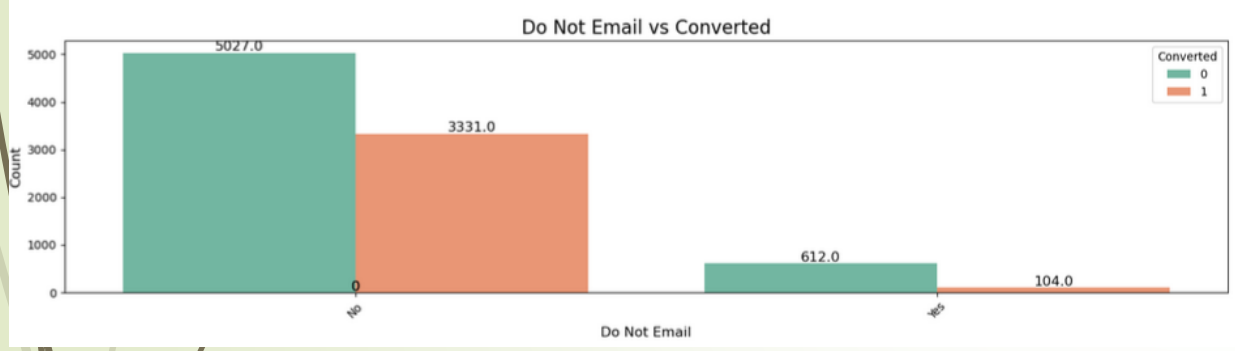
- When it comes to **Lead Origin** the count of converted is high for **Landing Page Submission** but as depicted from the graph **Lead Add Form** has proportionally higher converted leads.
- Similar pattern is observed in **Lead Source** where **Google** has the higher count of conversion but proportionally the conversion is much higher in **Reference** and **Welingak Website**.

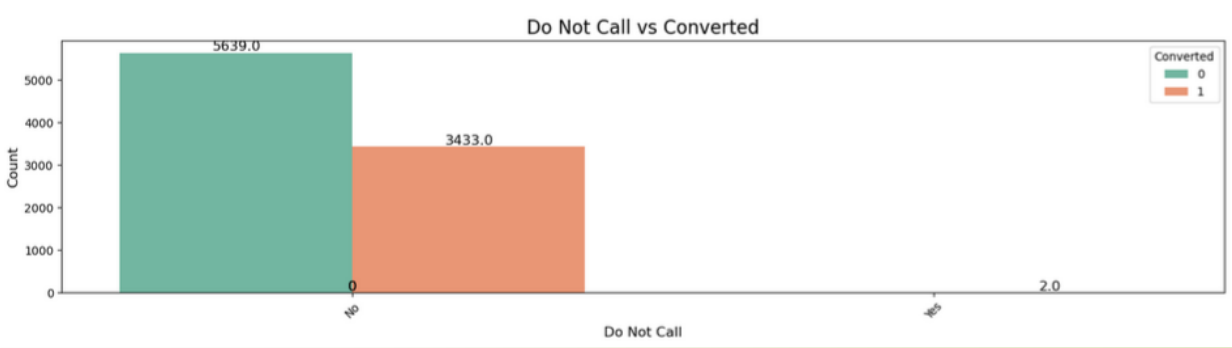




Exploratory Data Analysis

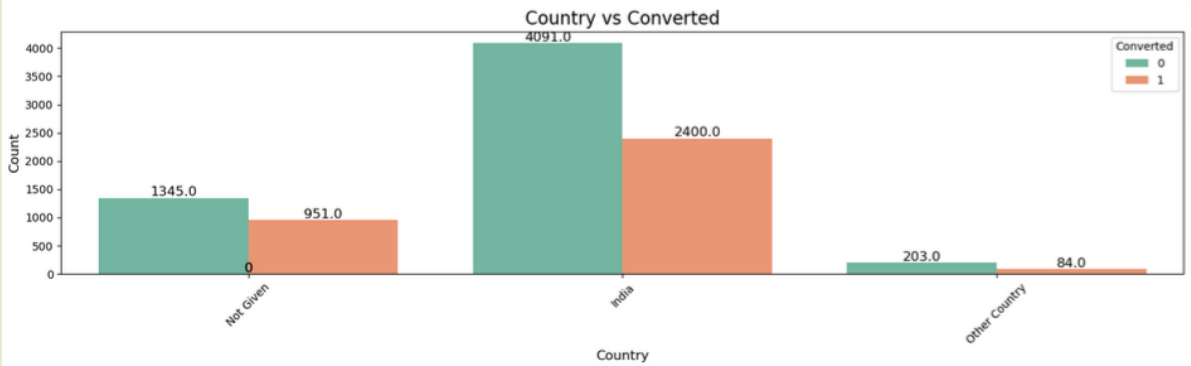
- Majority of the people opted for not being contacted via **Email** or **Call**. While the count of Conversion is similar in both the cases.

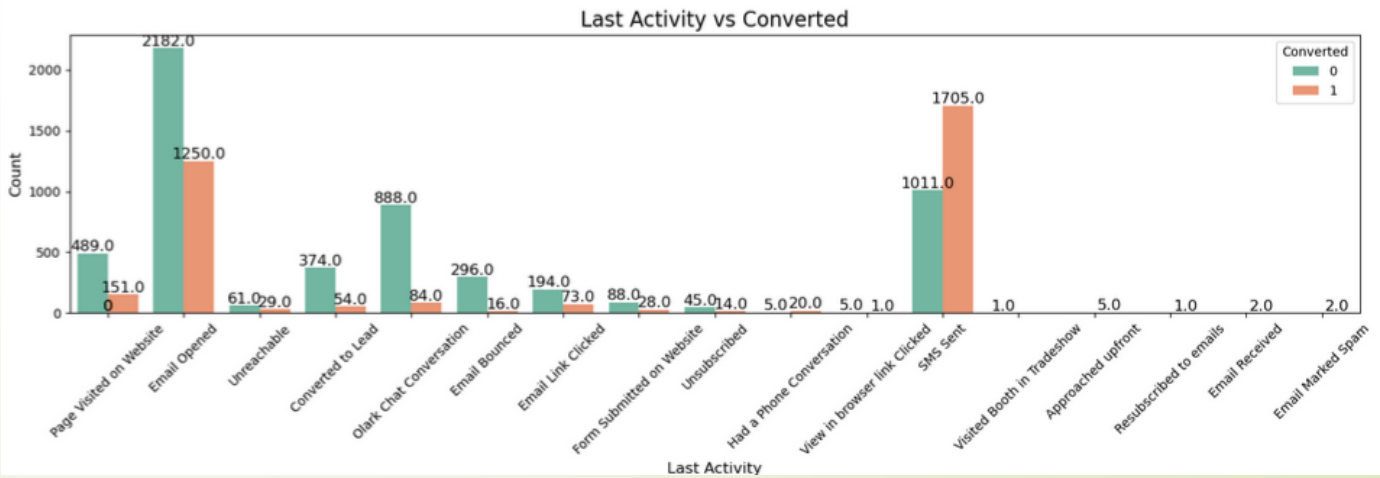




Exploratory Data Analysis

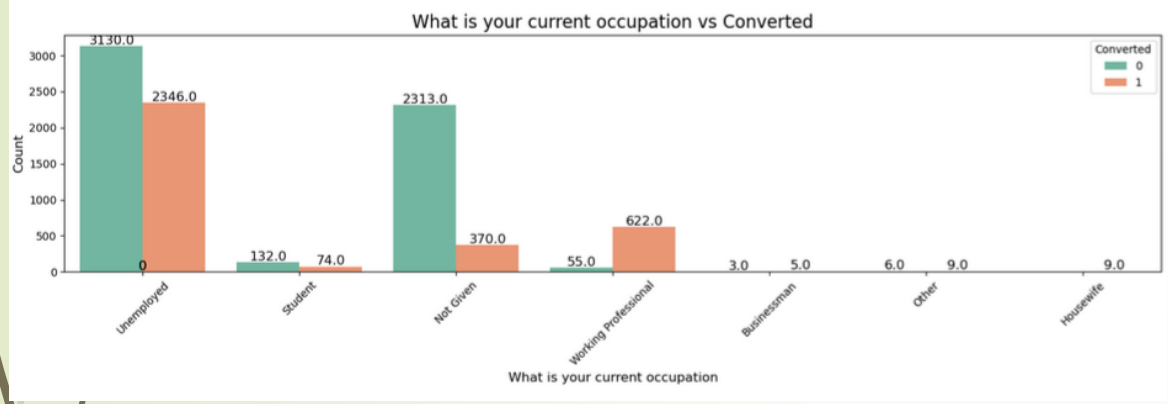
- While majority of the customers are from India, the **Converted Leads** count is **2400** which is also the highest followed by those who refrained from answering the demographics detail.
- An interesting observation is when the **Last activity** is **SMS Sent** the conversion is the highest with the count of **1705** along with a different proportion trend.

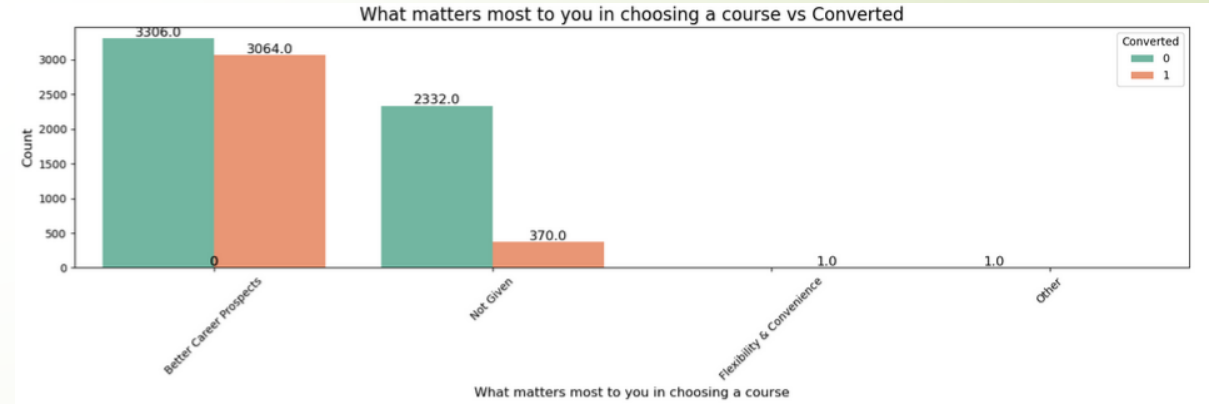




Exploratory Data Analysis

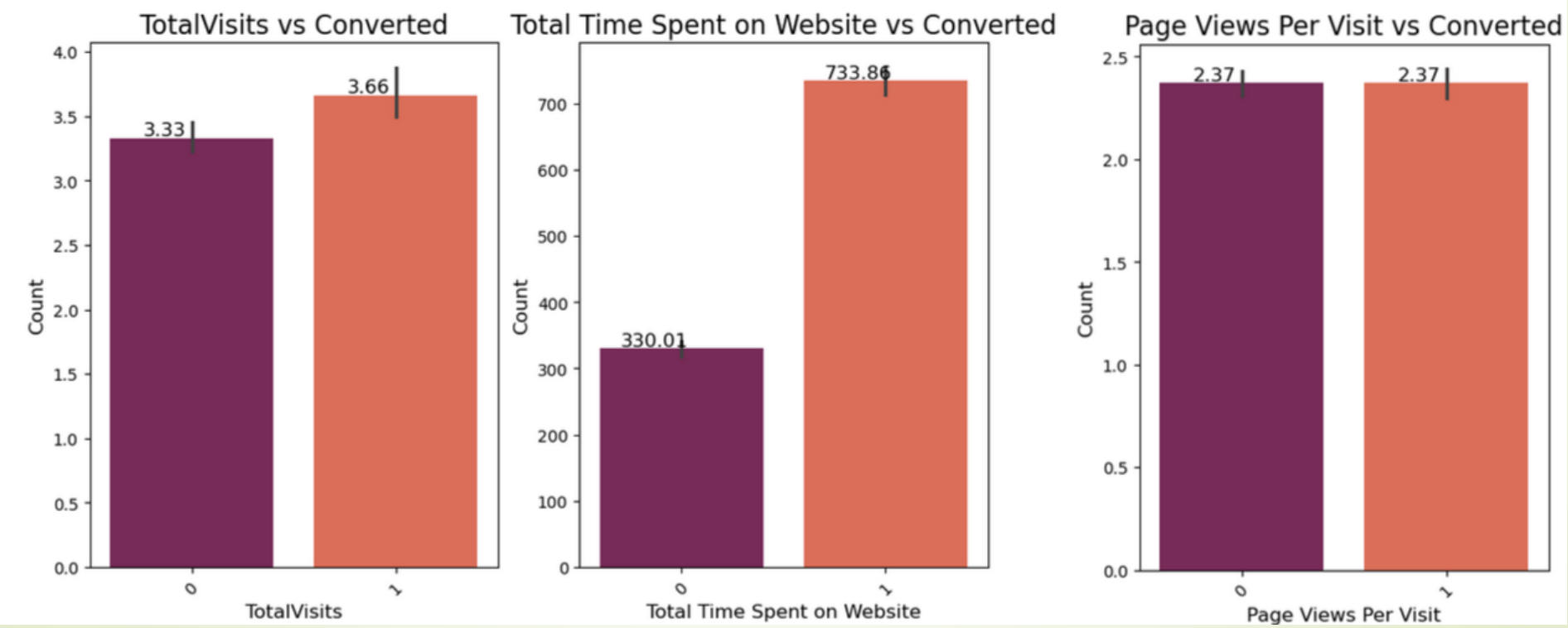
- **Unemployed** leads have the highest, **2346** converted leads. Although, the converted lead count is low the proportion pattern is quite the opposite with **Working Professional** with **622** leads converted while only **55** are not converted.
- Also, the count of conversion is highest at about **3064** for people whom **Better Career Prospects** is the reason for opting the course.





Exploratory Data Analysis

- **Total Visits, Total Time Spent on Website** have a comparatively greater count of Converted Lead at **3.66** and **733.86** respectively.
- **Page Views per Visit** has equal number of converted and non-converted leads.



Exploratory Data Analysis

- **Overall Insights: -**
  - **Engagement Variation:** There is considerable variation in user engagement (Total Visits, Time on Site, and Page Views per Visit). While some users interact minimally, a significant subset is highly engaged.
  - **Personalization Opportunity:** The varying behaviors (e.g., short visits vs. long visits, few pages vs. many pages) suggest opportunities for personalizing content or improving the user experience, especially for users who spend little time on the site or view only a few pages.
  - **Potential for Improvement:** There may be an opportunity to improve user retention for those who spend little time or view only one page by optimizing landing pages or offering engaging content to encourage deeper exploration of the site.
  - **There are no outliers present**

Feature Selection and Transformation

Feature Transformation

- **Categorical** columns needs to be converted to **Dummy** variables. The following are the categorical columns: -

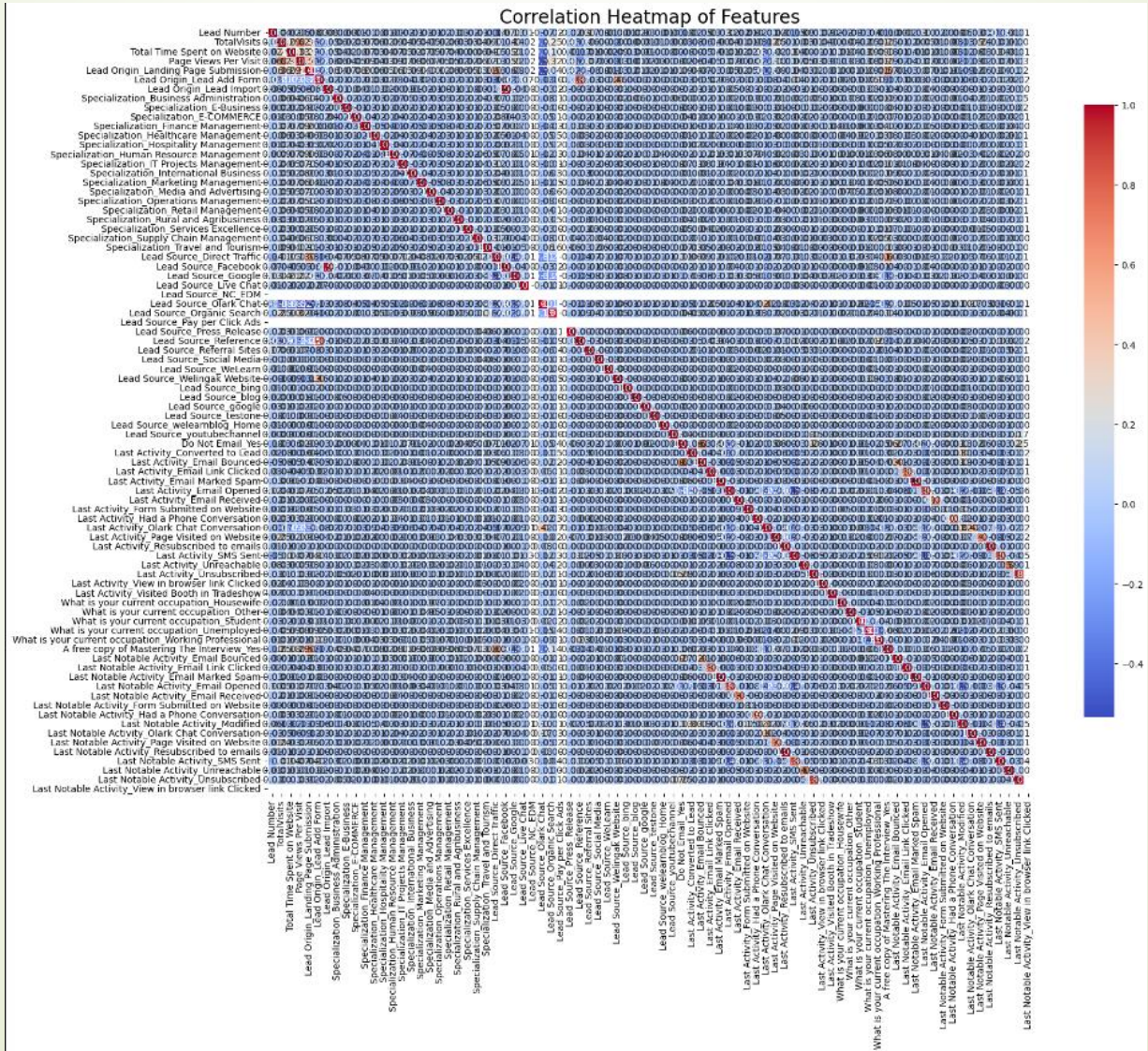
```
categorical_columns = lead3.loc[:, lead3.dtypes == 'object'].columns

print(categorical_columns)
```

```
Index(['Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call',
       'Last Activity', 'Country', 'Specialization',
       'What is your current occupation',
       'What matters most to you in choosing a course', 'Search',
       'Newspaper Article', 'X Education Forums', 'Newspaper',
       'Digital Advertisement', 'Through Recommendations',
       'A free copy of Mastering The Interview', 'Last Notable Activity'],
      dtype='object')
```

- **Numerical** columns needs to be scaled so that different value range are standardized or brought down to a common scale. Here we use **MinMaxScaler** to transform the numerical columns: -
  - **TotalVisits**
  - **Page Views Per Visit**
  - **Total Time Spent on Webstie**

Feature Transformation

- A **heatmap** is plotted depicting the **correlation** between features as shown below**: -**



Feature Selection

- A huge set of **82 features** are available and not every feature that is available at this point is exactly useful or contributing to the performance improvement of the model. Hence certain feature selection techniques mentioned below are employed in the case study: -
  - **Recursive Feature Selection(RFE) : -** An automated technique that can help us identify the most important features by iteratively removing the least important ones. We decided on keeping **15** most important features.
  - **Statistical significance of features : -** A manual technique that helps to check whether the variable is statistically as significant to training the model by checking the **p-value** for each of the variable. We found **x** features insignificant and thus those were dropped.
  - **Variance Inflation Factor(VIF) : -** Helps in detecting multi-collinearity among the feature set. **VIF > 5** is the threshold to drop the feature. There were **x** features dropped.

- Post Feature transformation and Selection, a **Logistic Regression** model was trained as it was one of the goals in the case study.

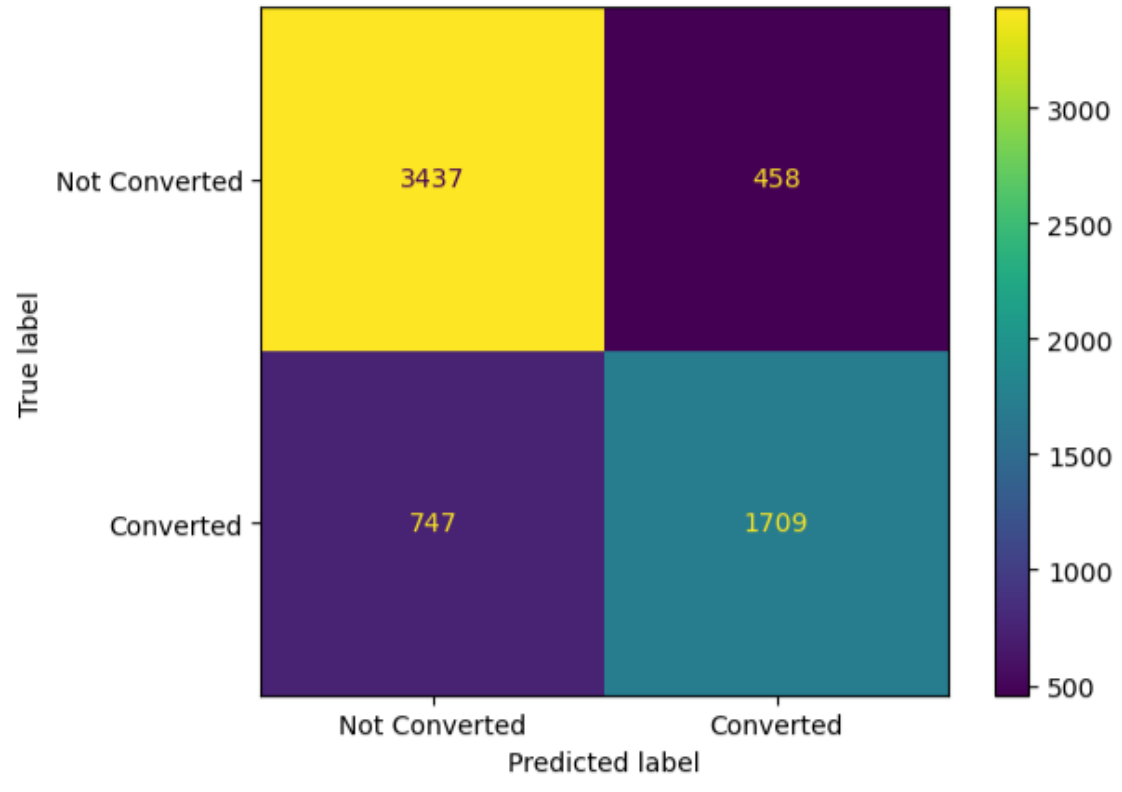
Model Evaluation

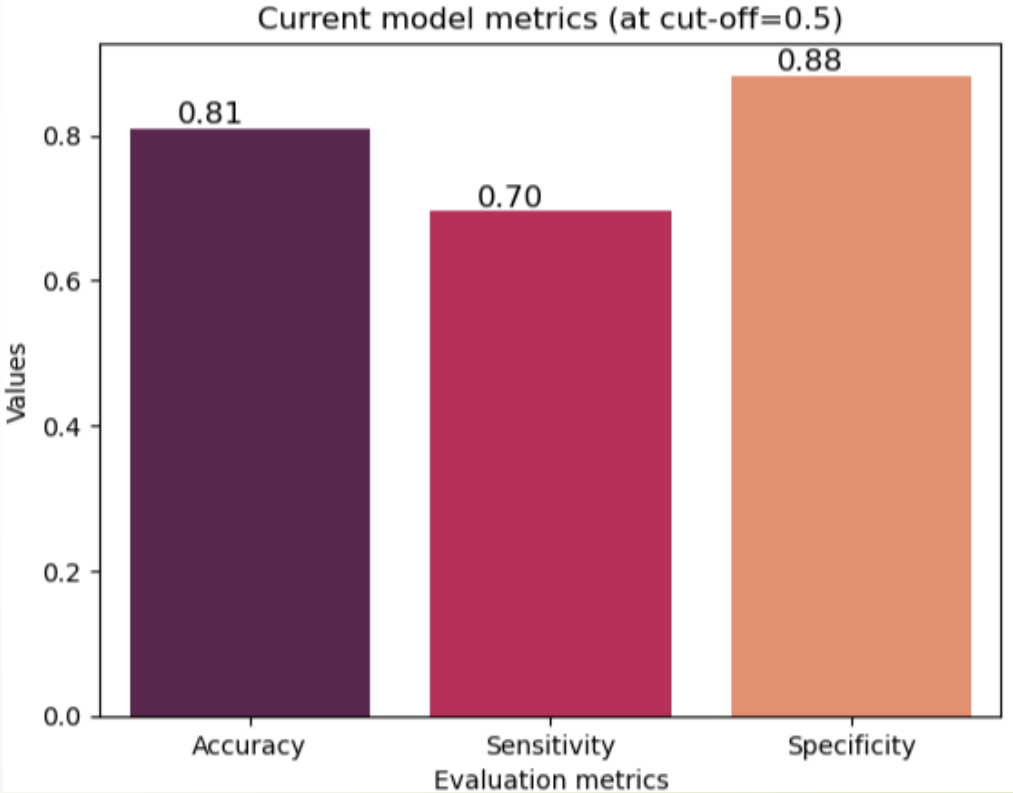
Model Evaluation – Train Set

- To evaluate the model the following parameters would be useful: -
  - Confusion Matrix
  - Sensitivity
  - Specificity
  - Accuracy
  - AUC-ROC curve
  - Accuracy VS Sensitivity VS Specificity and Precision-Recall curve to find optimal threshold for performance improvement.
- Out of all the mentioned metrics our main aim is to not miss on any of the lead rather to allow some of the leads to be left unattended. Hence, Sensitivity becomes one most important factor to consider followed by specificity and Accuracy
- Since we have imbalanced dataset we need AUC-ROC curve to check model's capabilities in identifying the leads. A higher score would indicate the model's classification capabilities are as expected.
- To further improve performance we need to find optimal thresholds which can be done by Accuracy VS Sensitivity VS Specificity chart and Precision-Recall curve.

Model Evaluation – Train Set

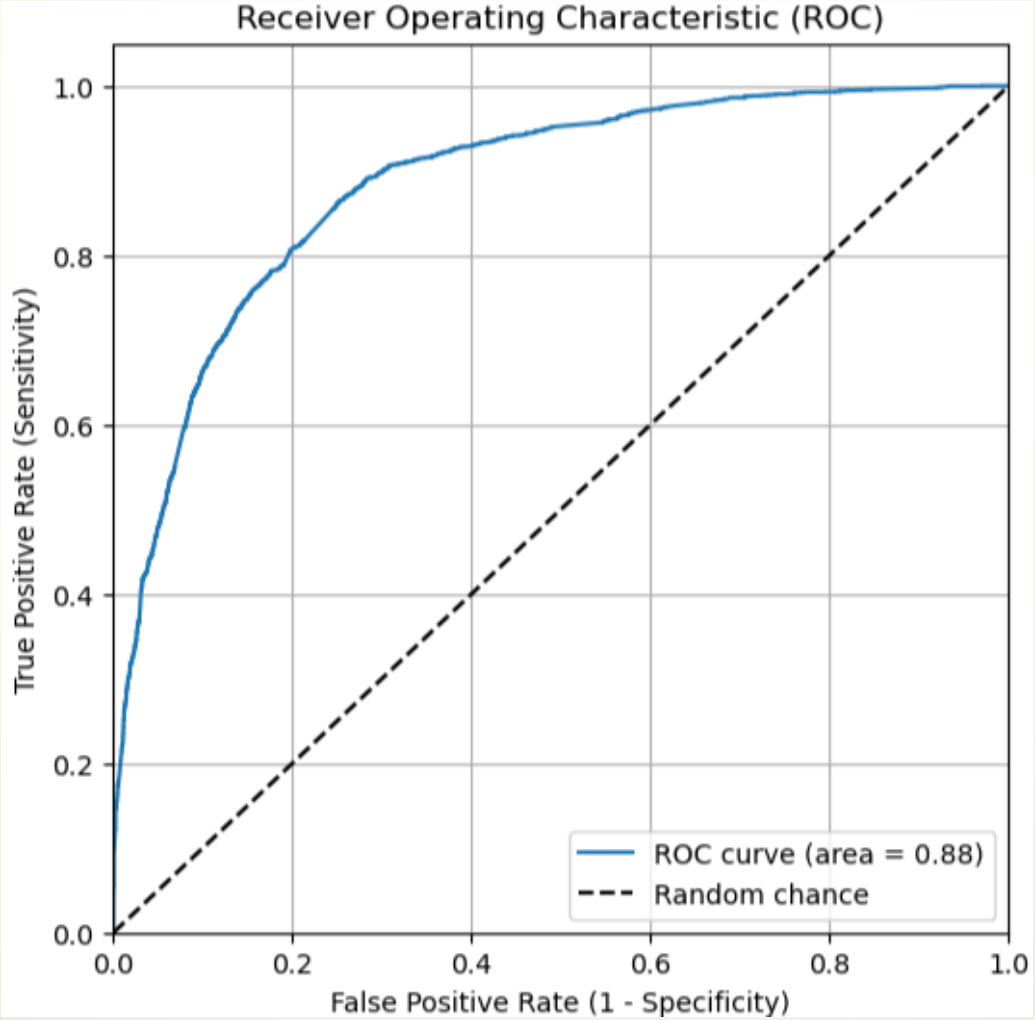
- As can be seen, the model in it's default configuration has predicted **1709** points as **Converted** while mistaking **747** points as **Not Converted**. The other performance metrics are as mentioned in the Model metrics chart





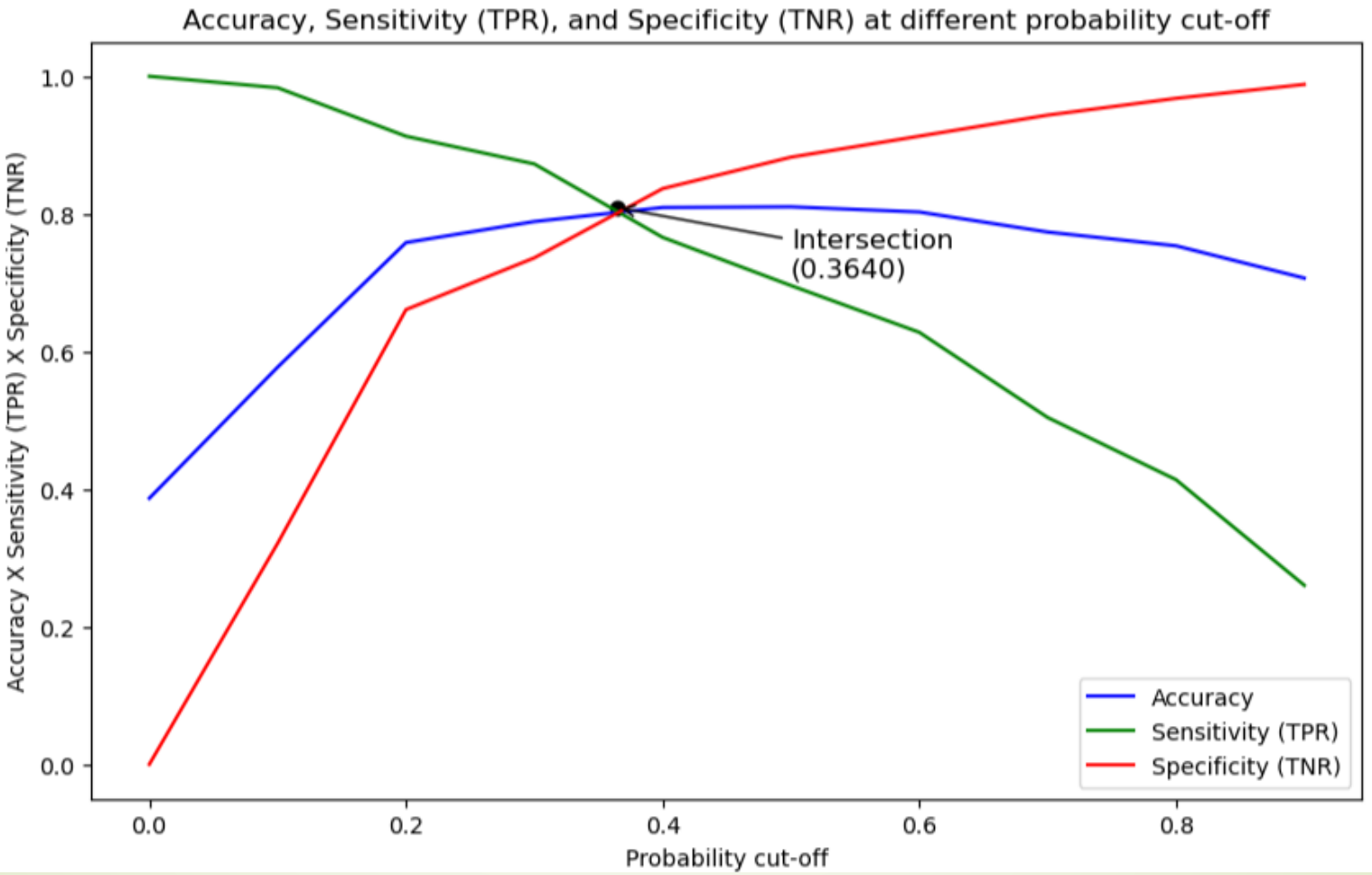
Model Evaluation – Train Set

- To check model's performance visually in understanding it's capability to distinguish we first tested with AUC-ROC curve. The model has a curve near the top left with a score of **0.88.**



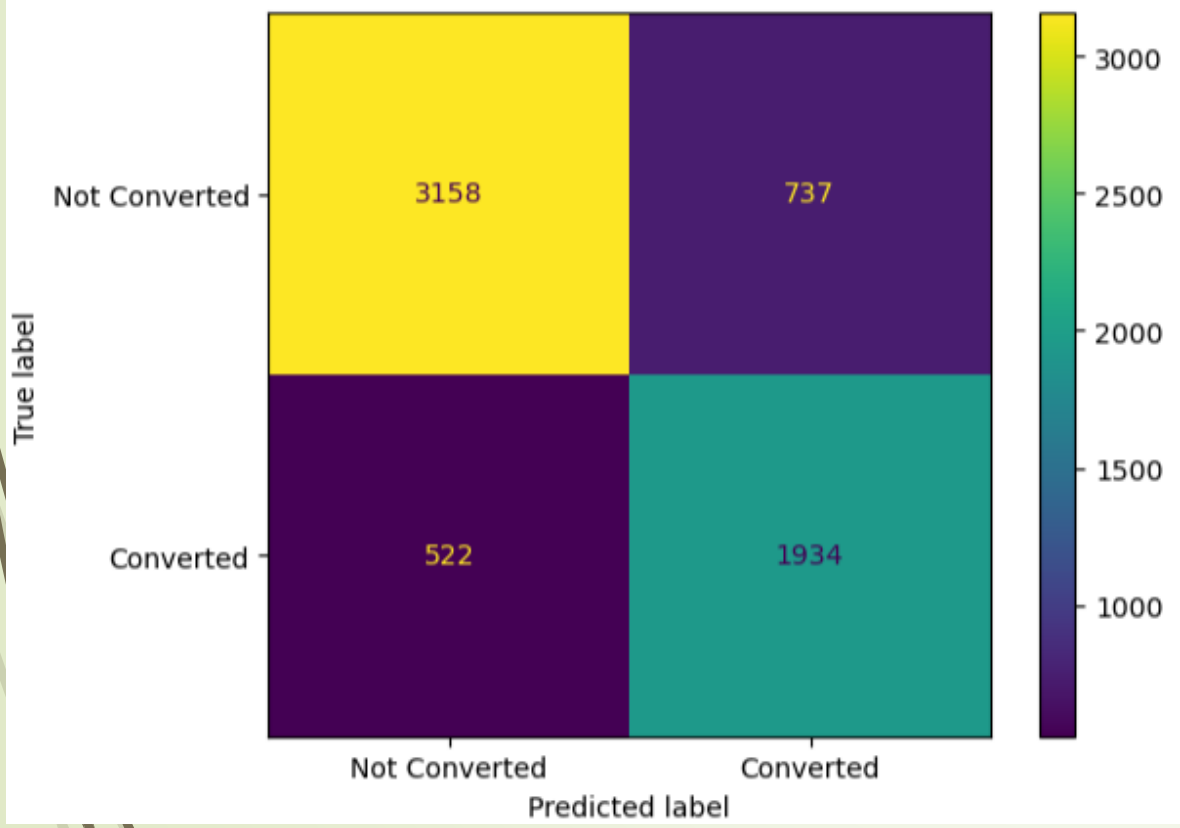
Model Evaluation – Train Set

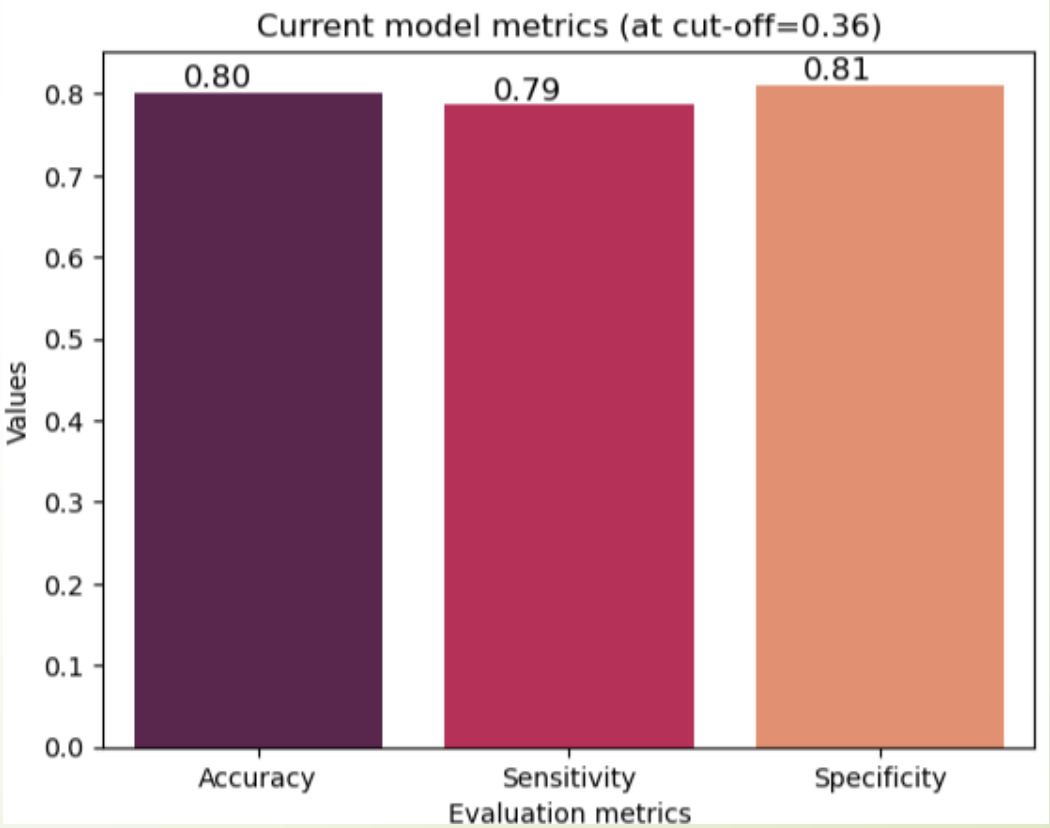
- In order to further improve the model, we decided to check for model's optimal threshold or cut-off point by plotting the graph shown below. A cut-off value of **0.36** is where all the points meet.



Model Evaluation – Train Set

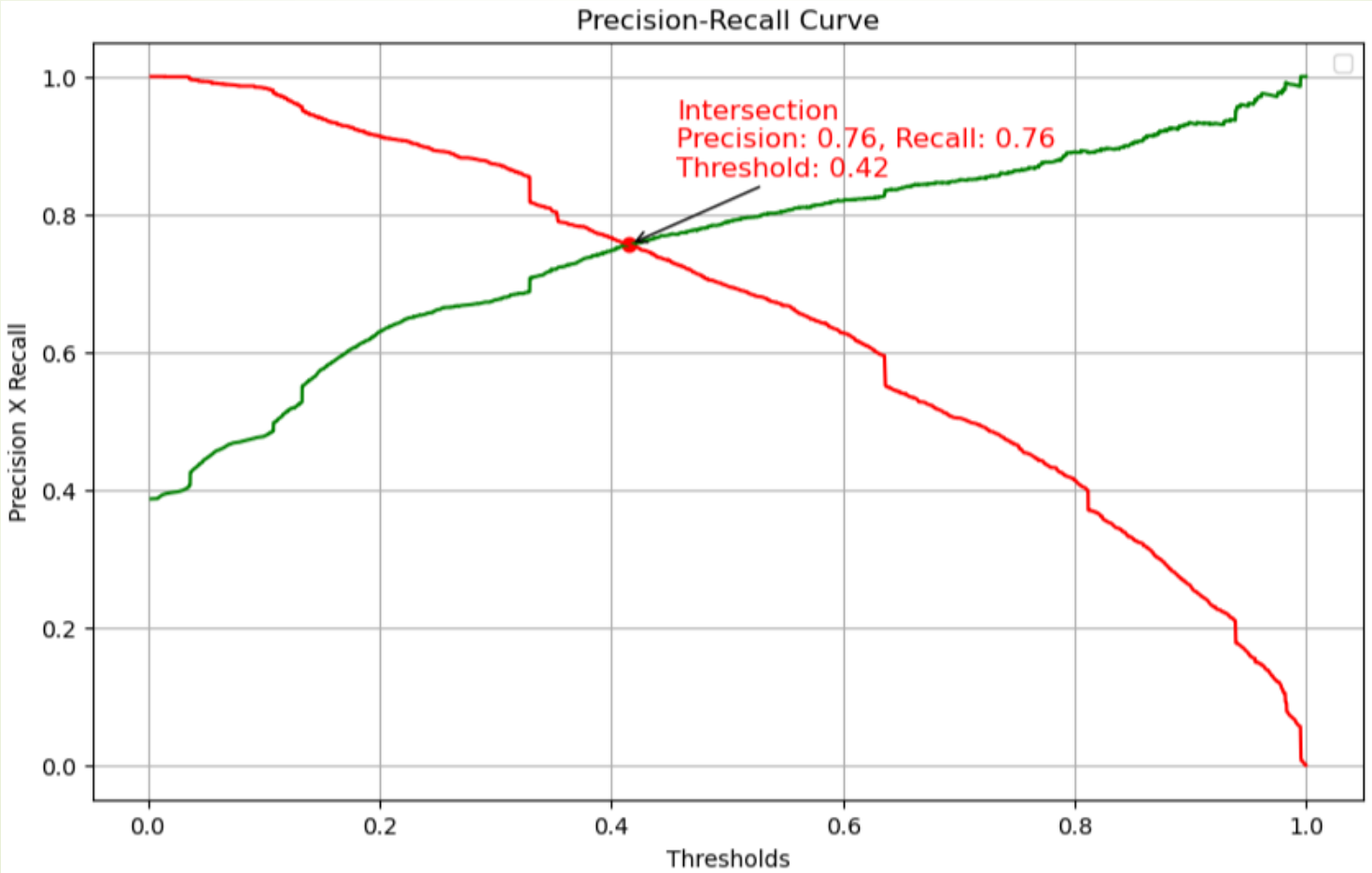
- As can be seen from the **Confusion Matrix** the **True Positives** have increased from **1709** to **1934** whereas the **False Negatives** have reduced from **747 to 522**.
- As per the **Evaluation Metrics** graph there is an increase in **Sensitivity** with reduction in **Specificity**.





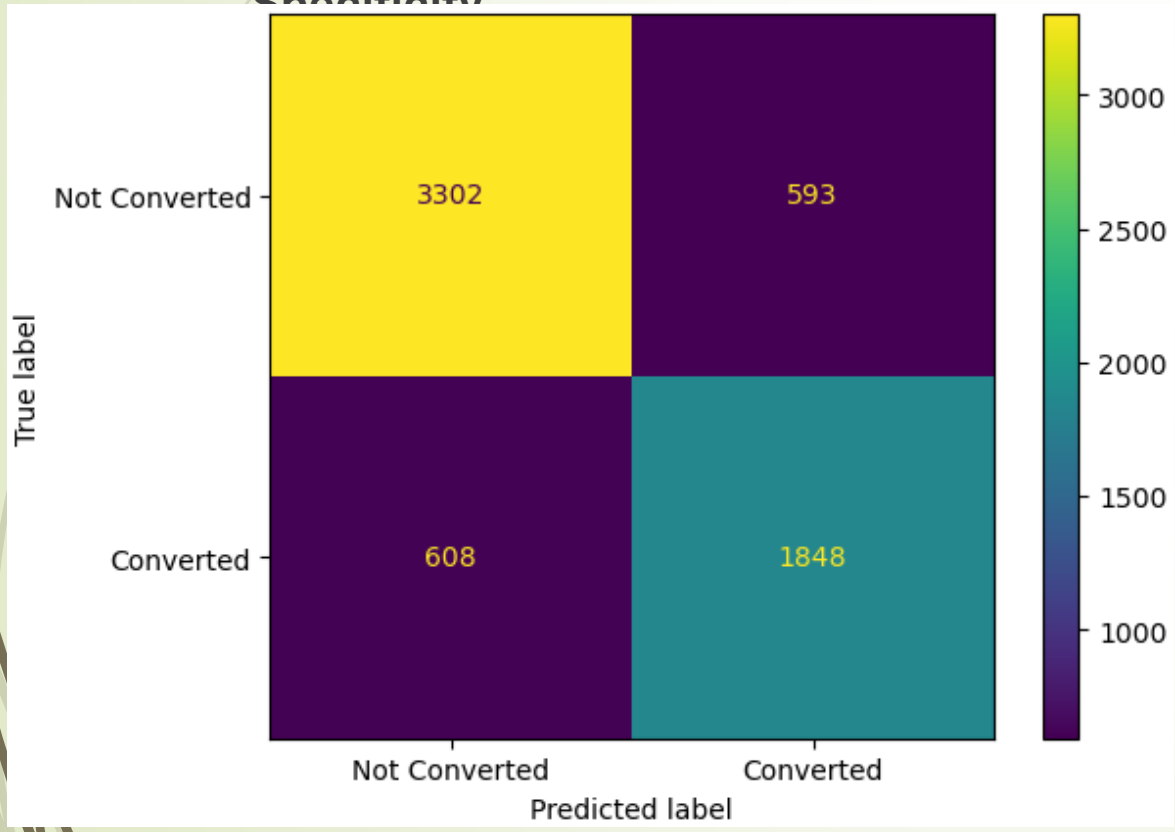
Model Evaluation – Train Set

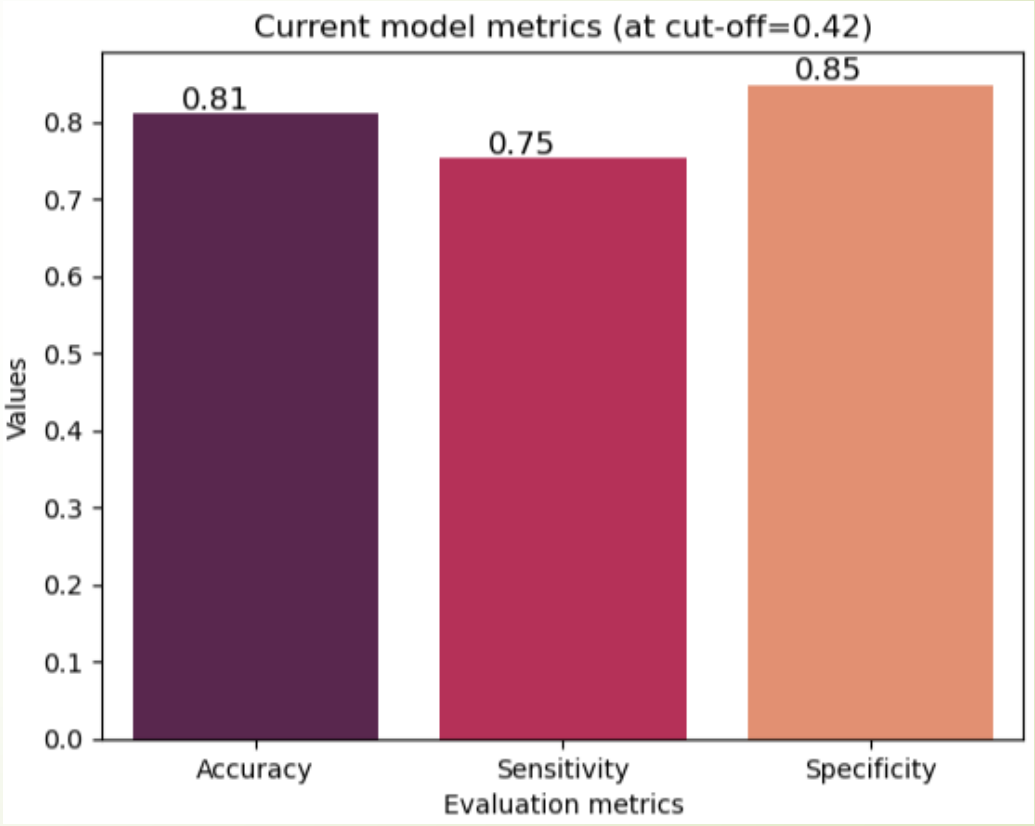
- In order to further improve the model, we decided to check for model's optimal threshold or cut-off point by plotting the graph shown below. A cut-off value of **0.42** is where all the points meet.



Model Evaluation – Train Set

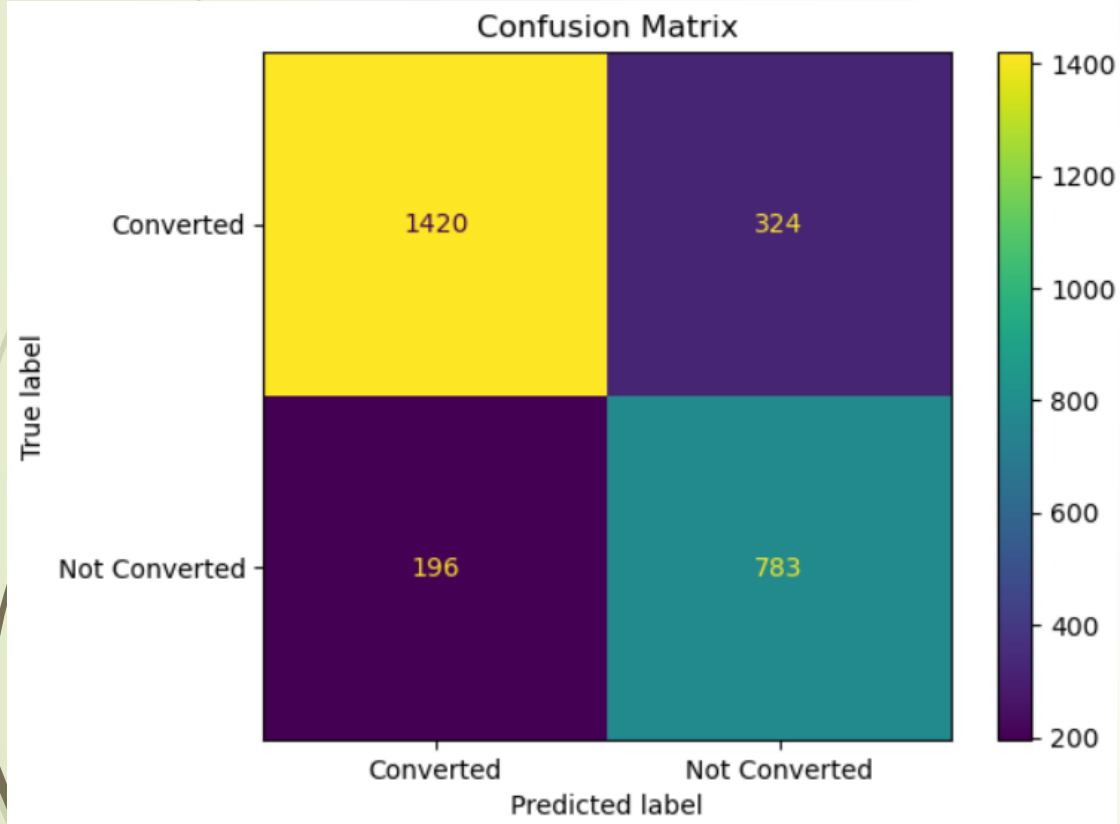
- As can be seen from the **Confusion Matrix** the **True Negatives** have increased from **3158** to **3302** whereas the **False Negatives** have reduced from **737 to 593**.
- As per the **Evaluation Metrics** graph there is an increase in **Sensitivity** with reduction in **Specificity**

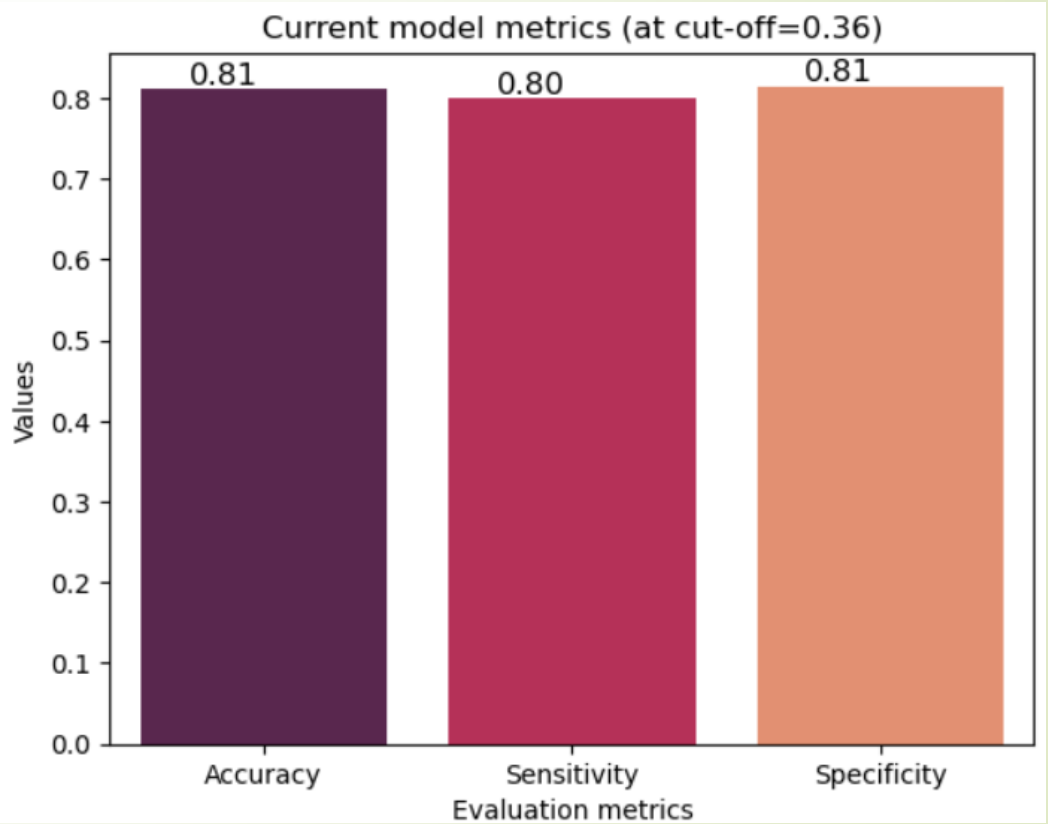




Model Evaluation – Test Set

- As per the **Evaluation Metrics** performed on the **Test dataset** graph the **Accuracy**, **Sensitivity** and **Specificity** are all at around **80%.**

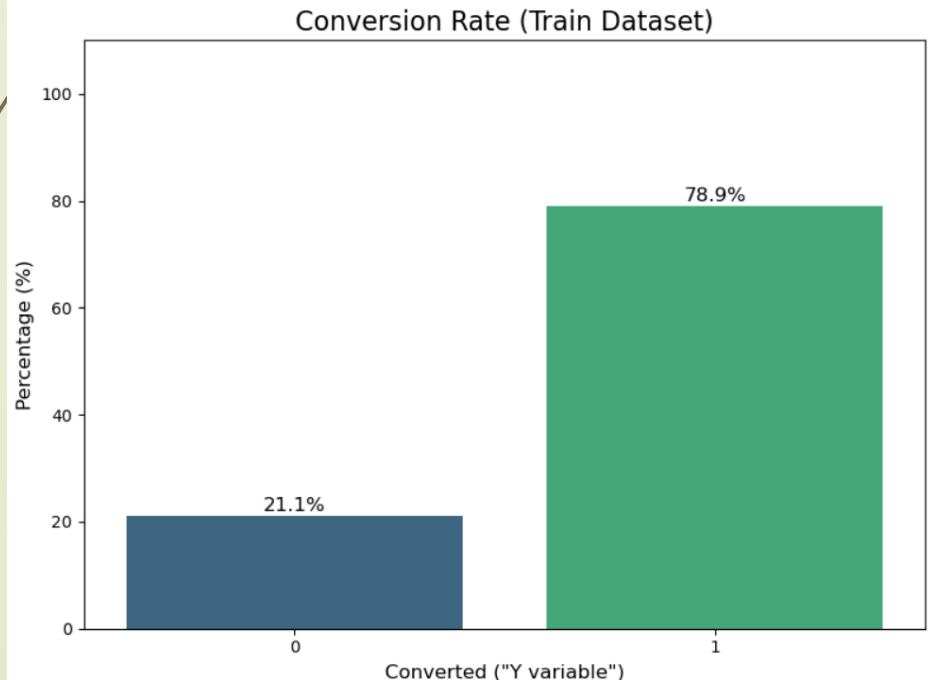


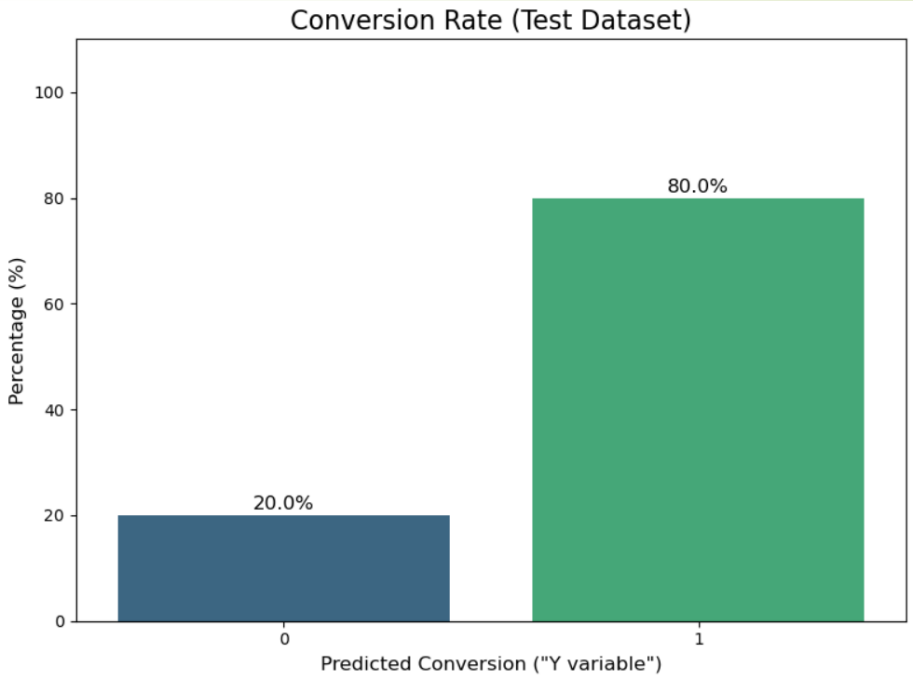


Business Results and Recommendation

Business Results

- The **CEO** of **X Education** provided a ballpark target **lead conversion rate** to be around **80%. Sensitivity** measures the proportion of actual positives correctly identified by the model. We have achieved a **Sensitivity** of **80%,** which in turn, also means that our model will be able to identify or predict the **Hot Leads** or achieve **lead conversion 80%** of the time.
- For now, it can be verified by checking out the model's predicted converted leads in comparison to the actual converted leads. As can be seen in the graph **80%** of the converted leads are predicted as converted in test dataset while approx. **79%** on the train dataset.





Recommendations: -

- A Lead Score of **greater than 36** means that there is an **80%** chance of it converting to customer.
- Out of all the observed parameters **Total Visits**, **Total Time Spent on website**, **Lead Origin_Lead Add Form** are the most important predictors. It correlates positively with the **Conversion of Lead** which indicates if the customer's engagement based on the above parameters if increased can improve the conversion chance.
- Also, **Working Professionals and Welingak website** are major sources of leads and hence the presence should be improved for **Working Professionals** and on **Welingak website.**
- Company should also keep track of the changes In user behavior and periodically see how the model's performance has changed over time and incorporate them to improve the model.

THANK YOU