# LEAD SCORING CASE STUDY

Created By:
Aniket Patra
Dhaval Kale
Shivin Singh

# GUIDE:

- Problem Statement
- Objective
- Strategy Approach
- Data Preparation & Data Standardization
- EDA (Univariate)
- EDA (Bivariate)
- Correlation Test
- Model Building & Evaluation
- Insights
- Conclusion
- Recommendations

# PROBLEM STATEMENT:

- 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. The typical lead conversion rate at X education is around 30%,  its lead conversion rate is very poor.

- The company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

# OBJECTIVE:

- Assist X Education in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- The company has a requirement of building 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

# STRATEGY APPROACH:

- Importing the required libraries.
- Importing the data
- Cleaning data (removing null values, unwanted columns, etc).
- EDA on data for further understanding.
- Data preparation for model building.
- Building a Logistic Regression model.
- Assign a lead score to each lead.
- Testing the model over a training dataset.
- Evaluation of the model based on different measures and metrics.
- Test the model on the test dataset.
- Evaluation of the accuracy of the model based on different measures and metrics.

# DATA PREPARATION:

- Within the dataset the categorical variables that are having value as “Select” shows that no option was chosen and is hence it is treated as a null value. The value “Select” has been replaced with NaN considering the same.
- In total 7 columns with null values greater than 40% in the dataset and hence it is dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Dropped the columns that don't add any insight or value to the study objective (tags, country)
- Imputation used for some categorical variables and few 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 STANDARDIZATION AND OUTLIER CHECK

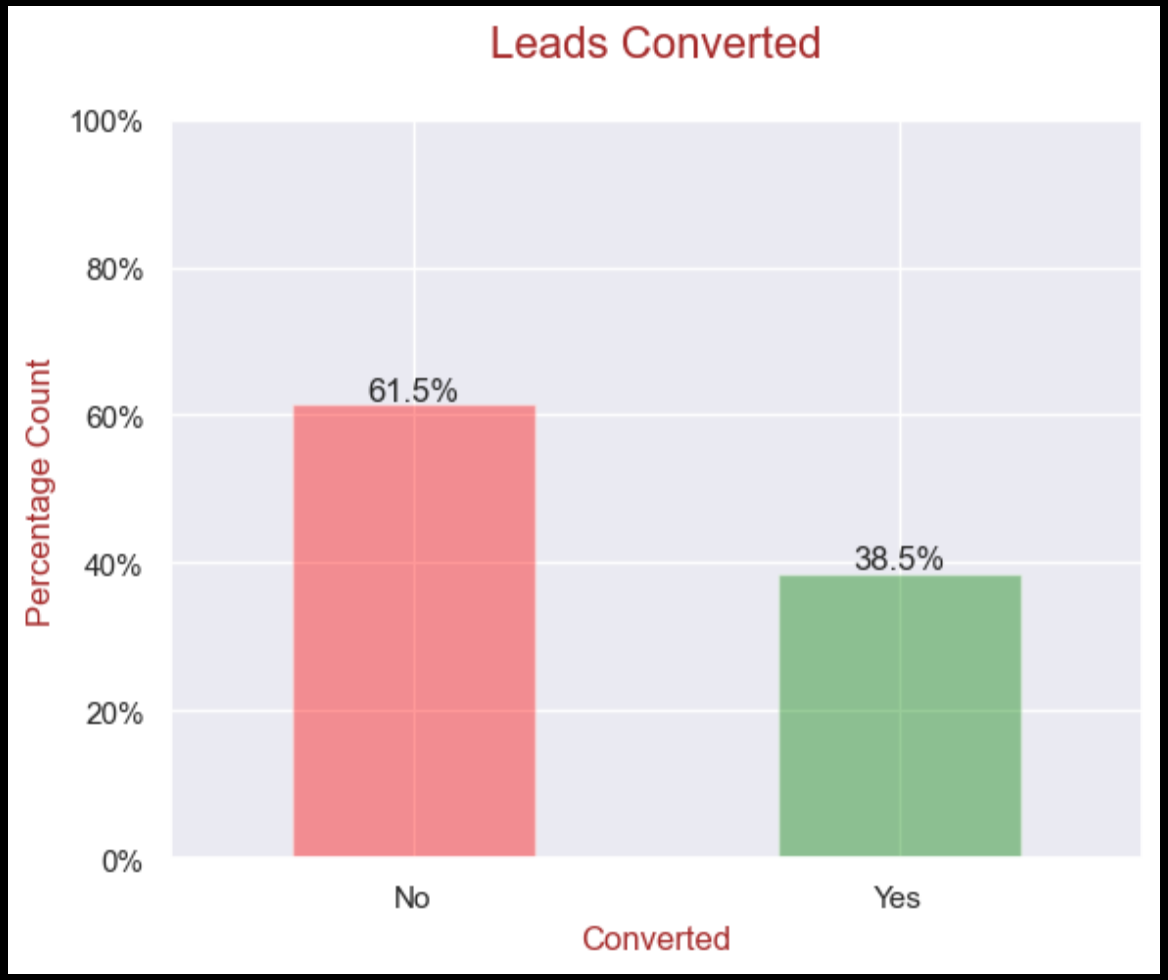
- Post analyzing the data it was noted that the "Lead Score" and "Last Activity“ columns have a number of labels whose value count is negligible. Hence, can be grouped together under "Others" to remove columns that are irrelevant for regression analysis.

- "Free_copy" & "Do Not Email" columns were both binary categorical columns. To standardize these, the values ‘yes’ and ‘no’ converted to 0 & 1.

- Columns having same value but are present in different cases for example "Google" & "google" are the same in "Lead Source", and are sanitized by replacing lower case values with upper case, which resulted in standardizing the data.

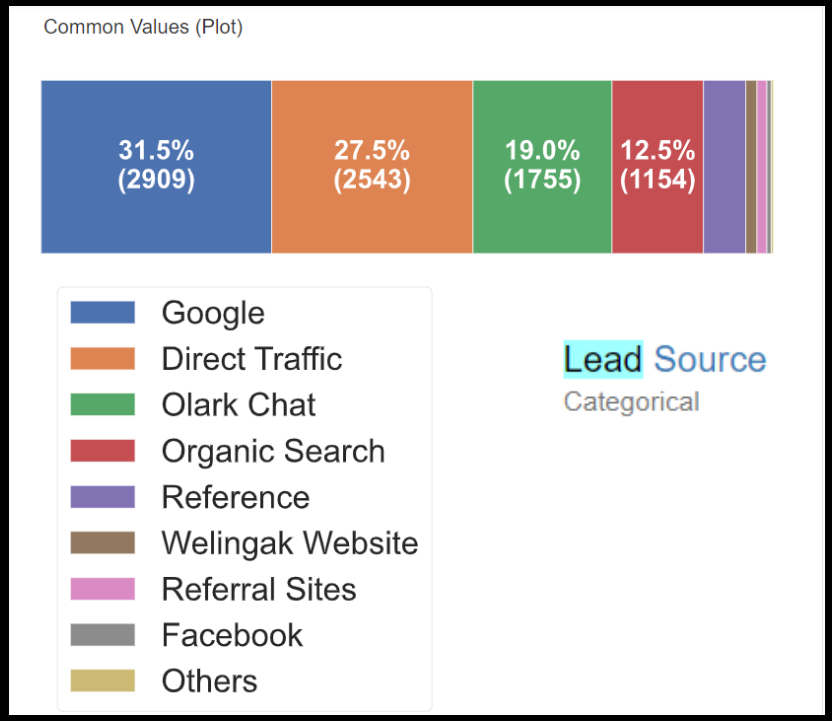
- An Outliers in the data were checked using boxplots and were treated by defining the upper and lower limits, replacing the values that lie outside the defined range with the correct value to remove an outlier.

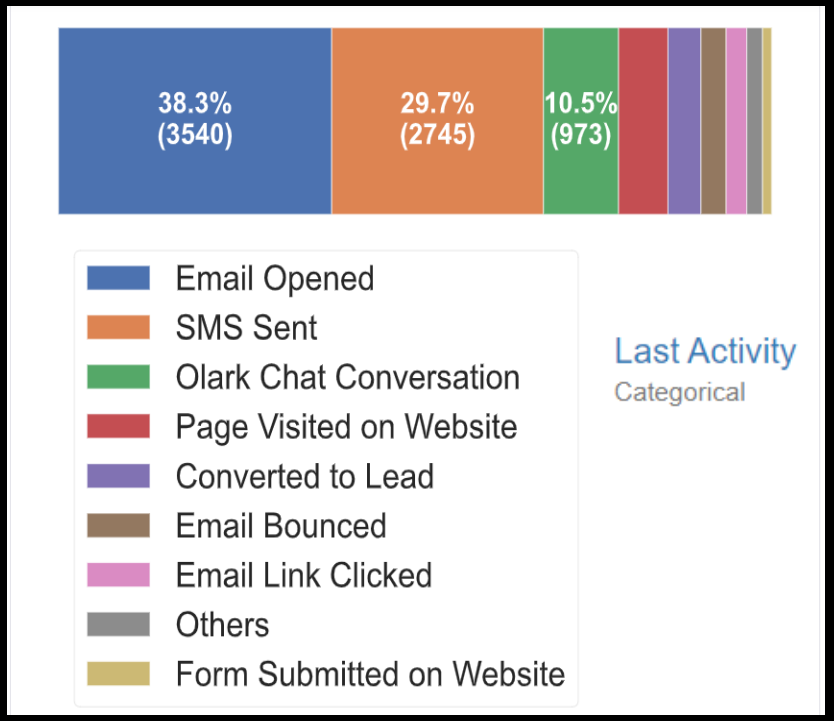
# DATA IMBALANCE




- 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)

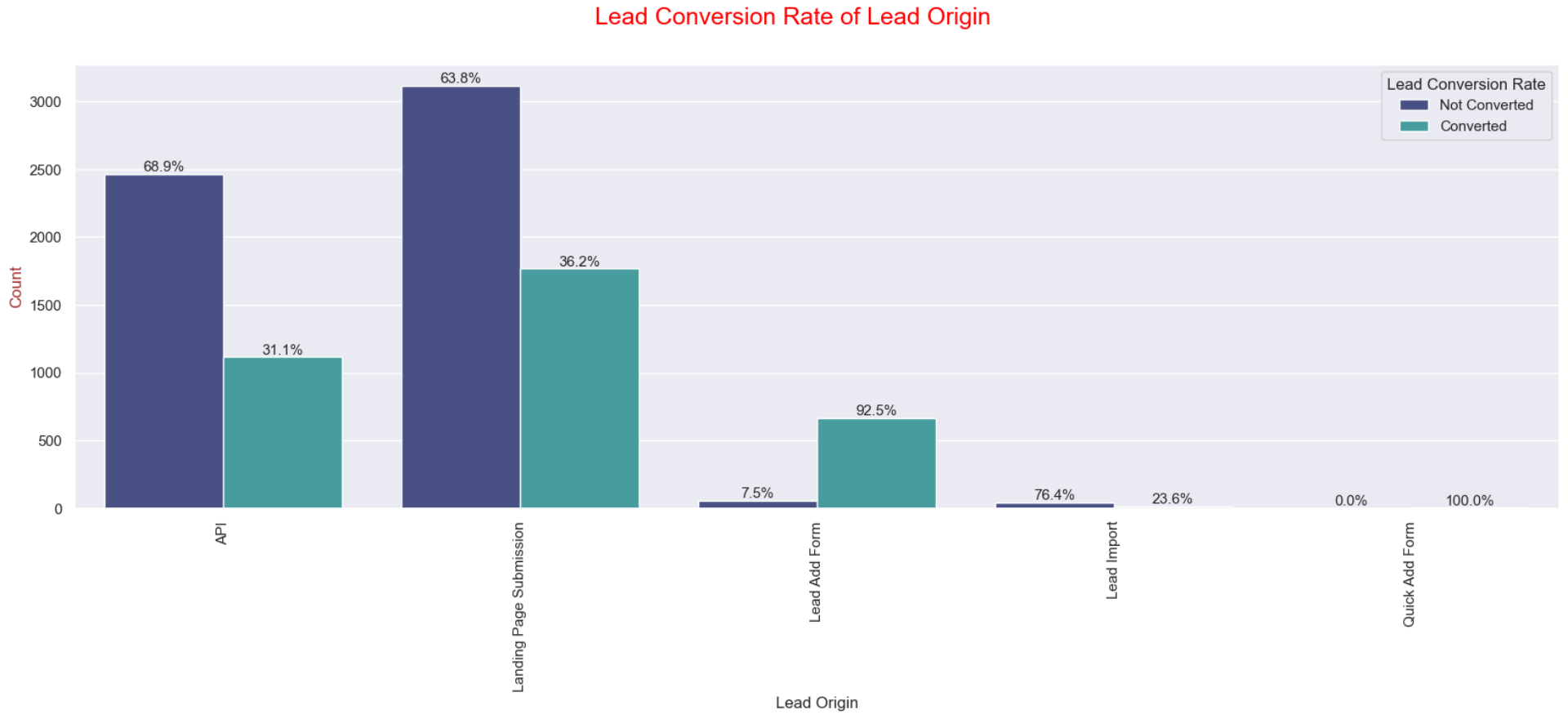
# EDA (UNIVARIATE ANALYSIS)




- Lead Source: 58% Lead source is from Google & Direct Traffic Combined.
- Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities

# EDA (BIVARIATE ANALYSIS)



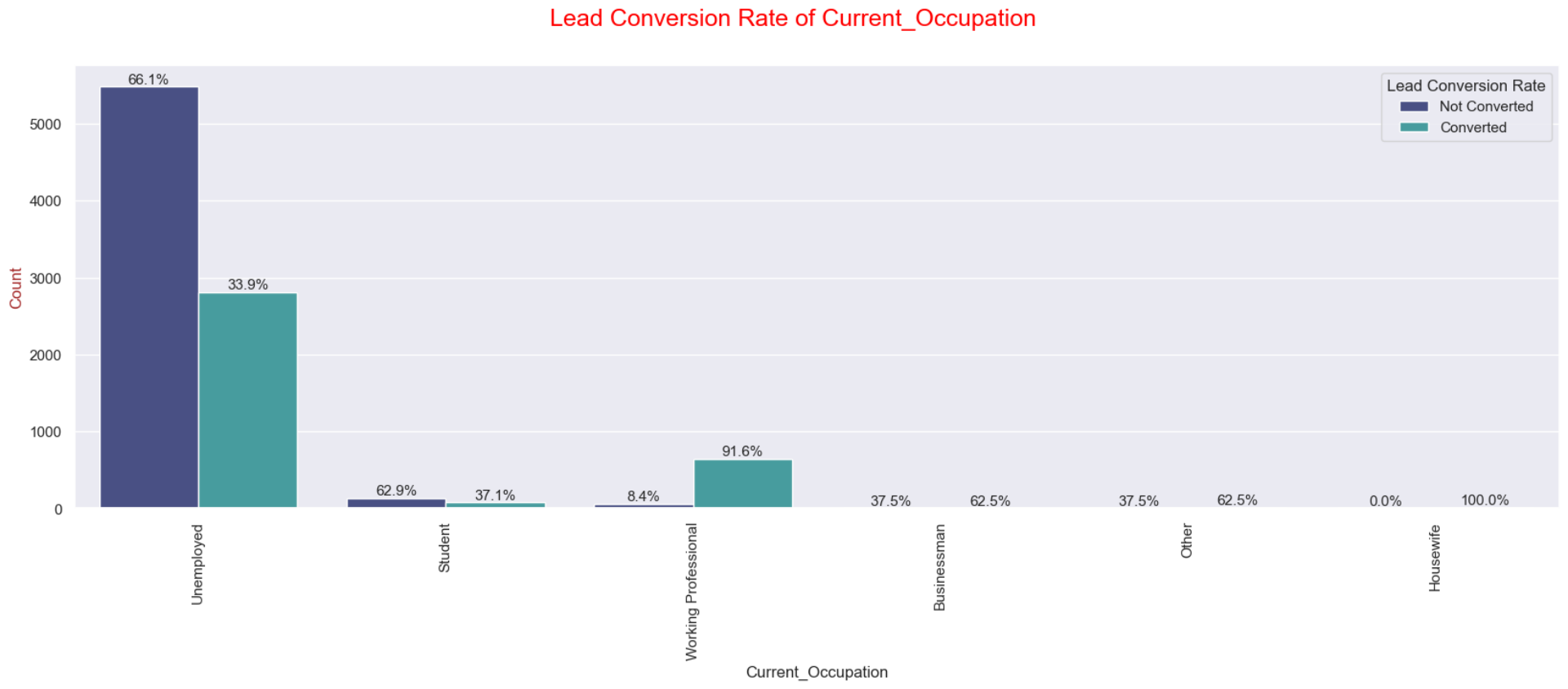

Conclusion from above graph:

(Lead Origin vs Conversion Rate)

Lead Origin:

- The majority of customers, 52.9%, which were identified through 'Landing Page Submission'.
- The lead origin, have a LCR of 36.2% followed by 'API' at 38.7%, with an LCR of 31.1%.

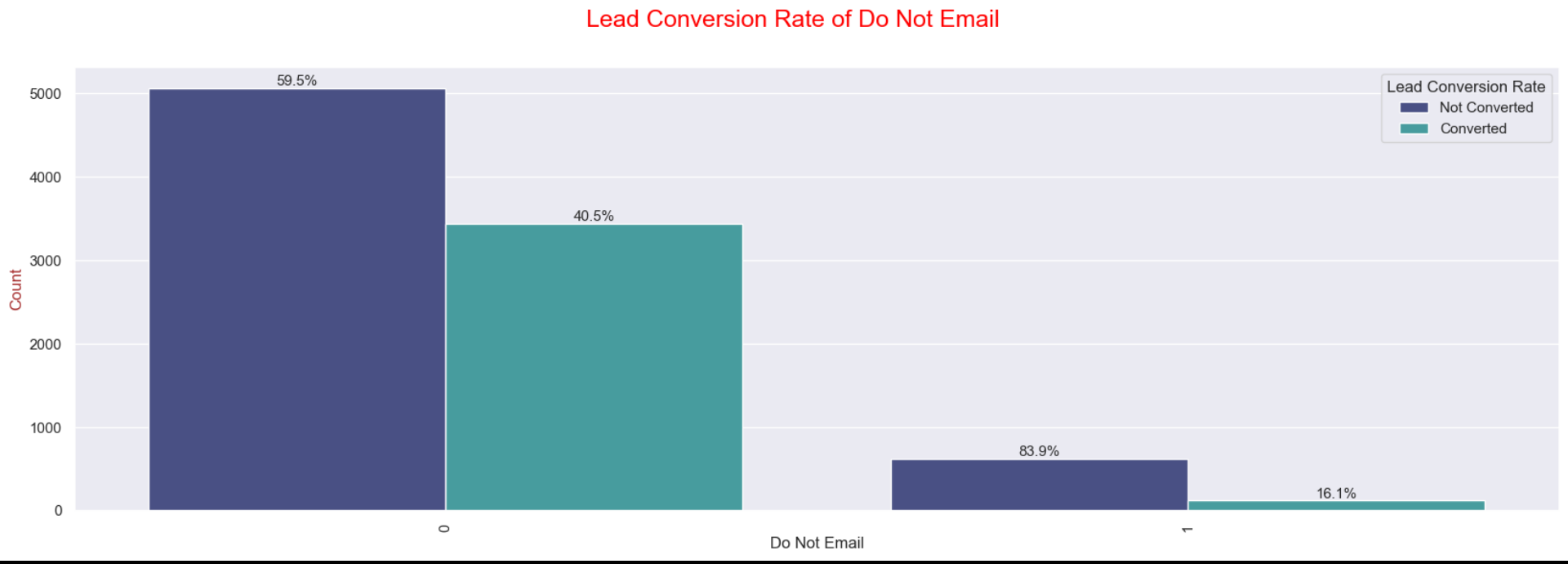
# CONTD:




(Current Occupation vs Conversion Rate)

- A significant proportion of customers, 89.7%, are unemployed but have a smaller LCR of 34% when compared to Working Professionals which are only 7.6% of the total customers but have an LCR of 92%.

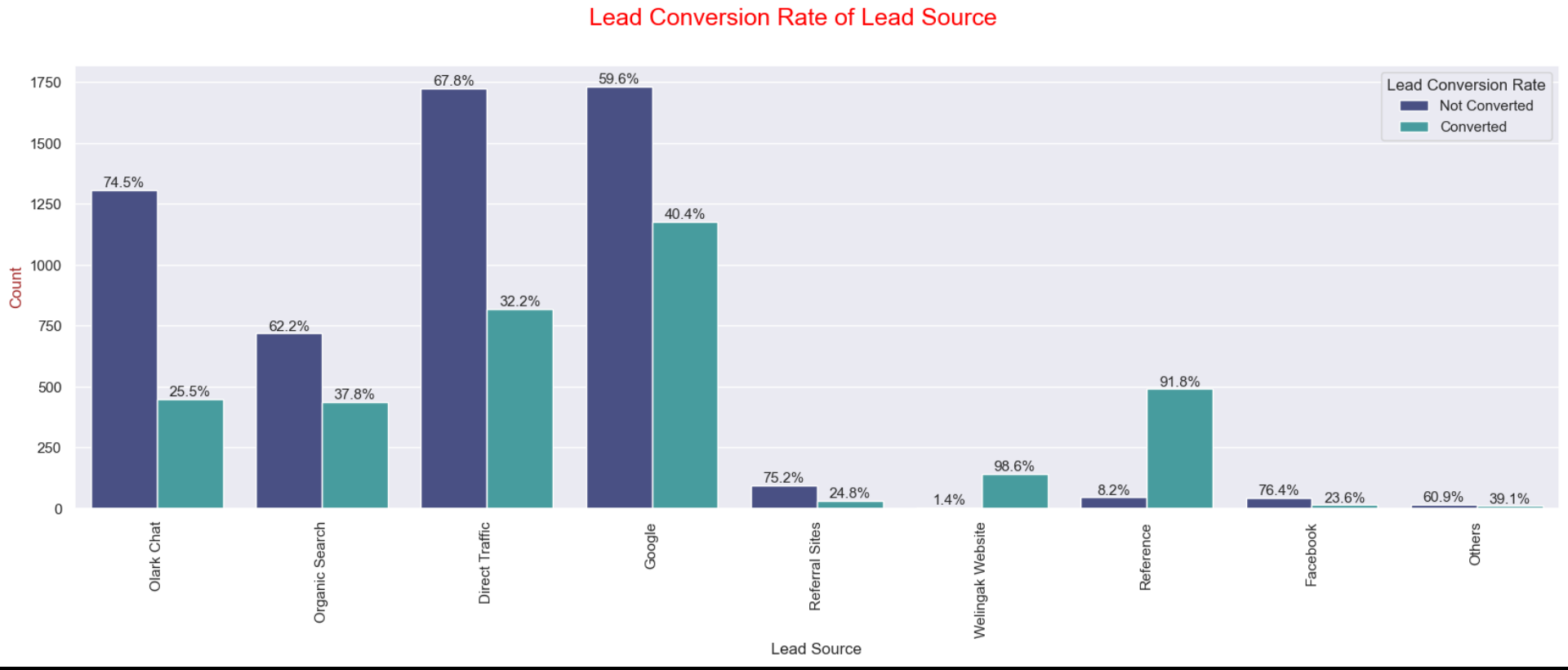
# CONTD:




(Do Not Email vs Conversion Rate)

- A large proportion of customers, 92.1%, do not want to receive emails about the course but have an LCR of 40.5%.

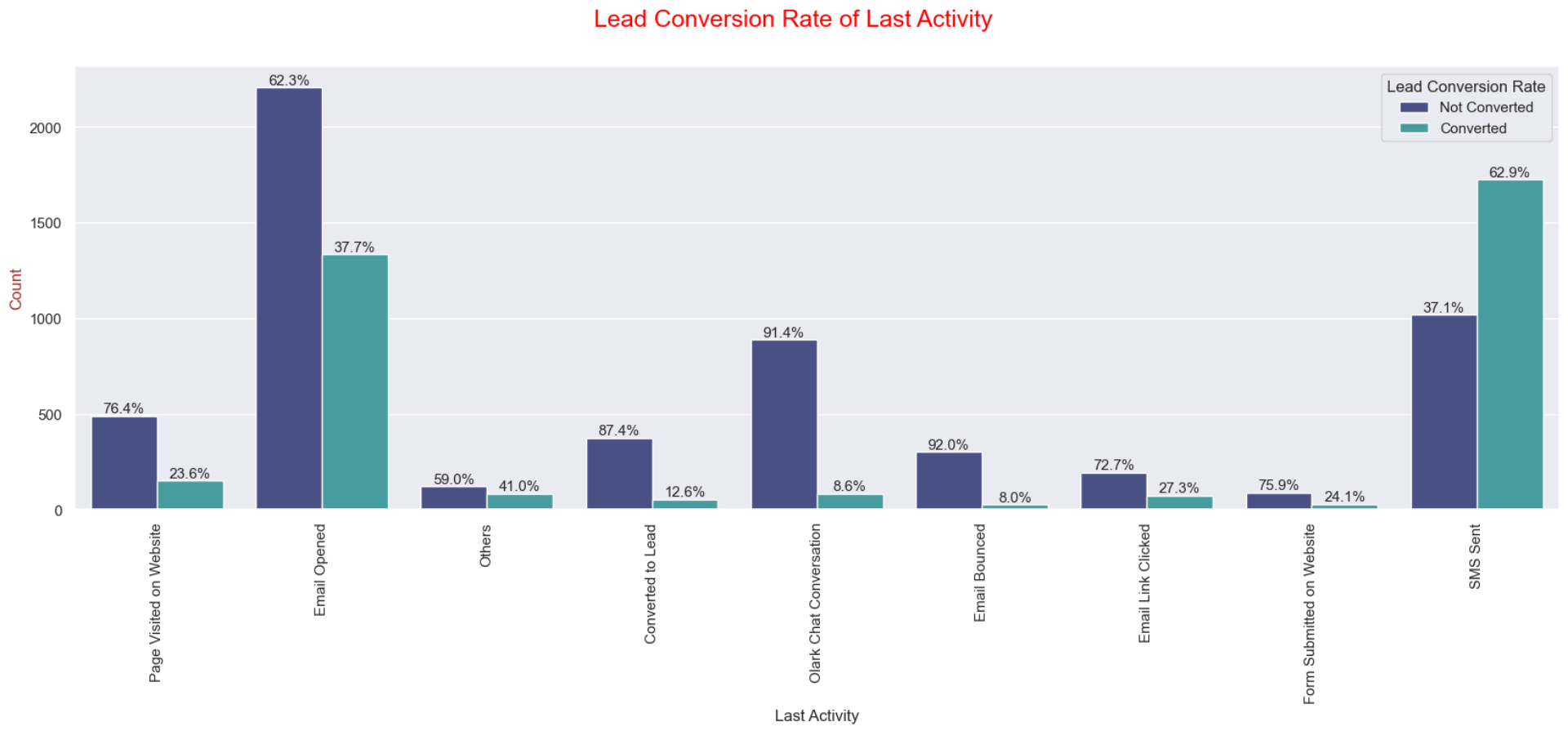
# CONTD:




## (Lead Source vs Conversion Rate)

- Google is the most effective Lead Source with an LCR of 40.4%, followed by Direct Traffic at 32.2% and Organic Search at 37.8% (contributing to only 12.5% of customers). Reference has the highest LCR at 91.8%, but there are only 5.8% of customers through this Lead Source.

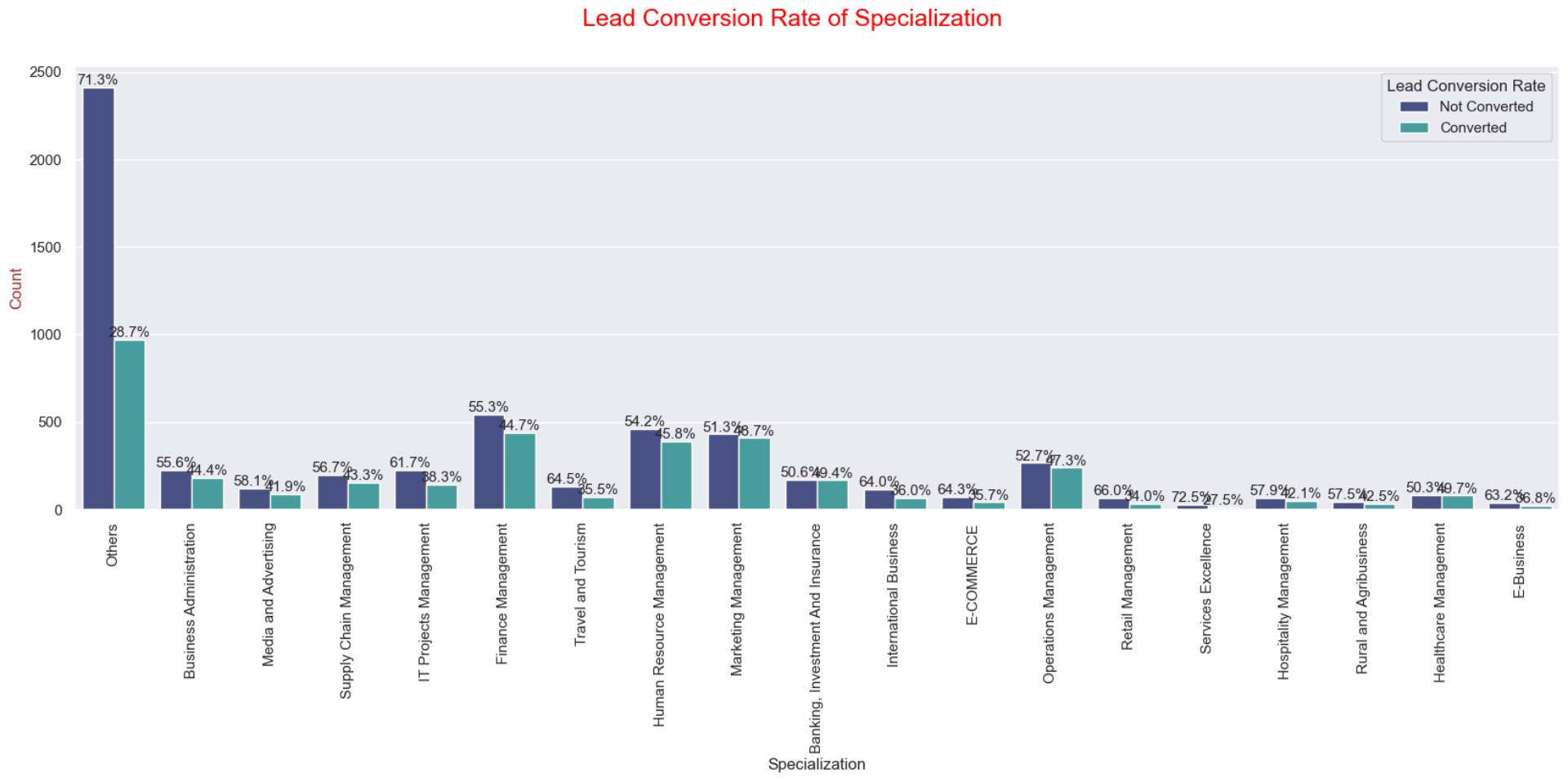
# CONTD:




(Last Activity vs Conversion Rate)

- SMS Sent and Email Opened are the most effective Last Activity types with `LCRs` of 62.9% and 37.7% respectively.

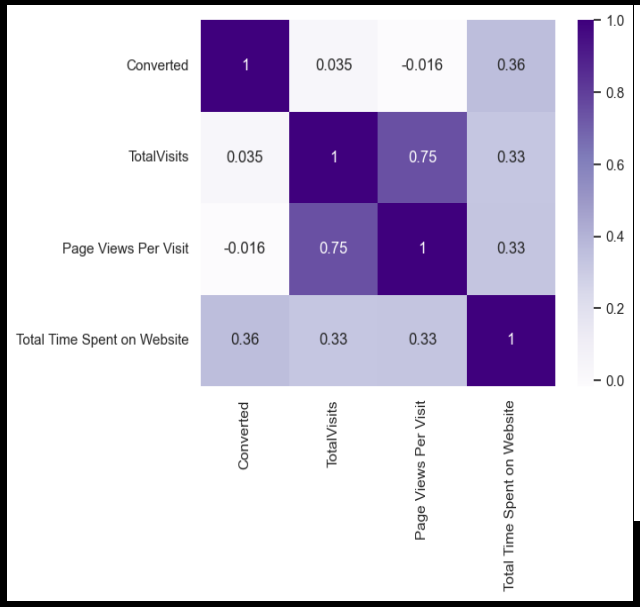
# CONTD:

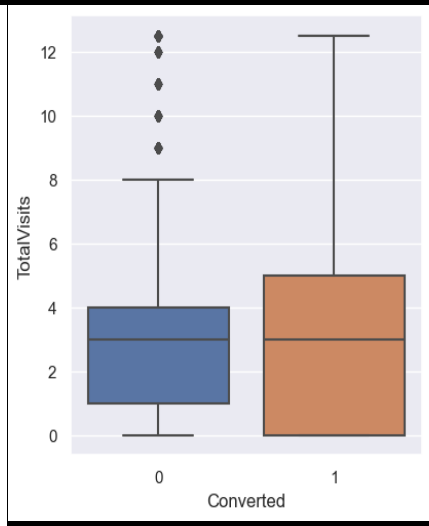


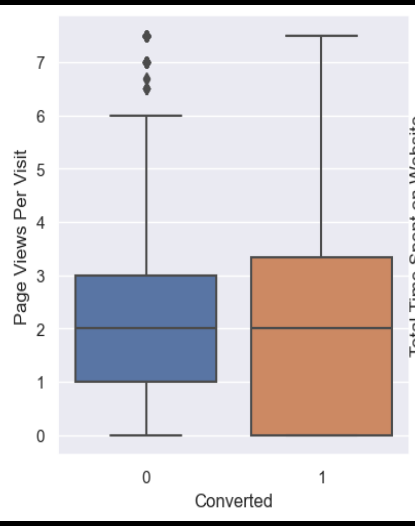

(Specialization vs Conversion Rate)

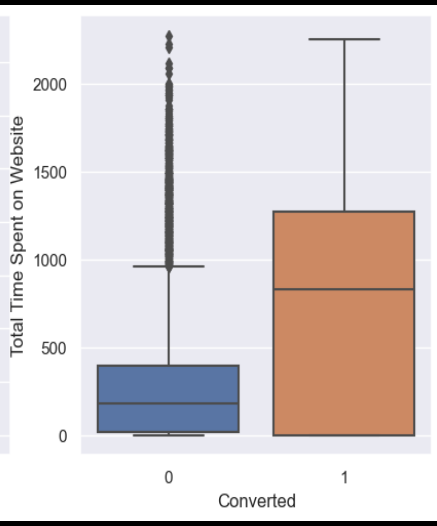
- Marketing Management, HR Management, Finance Management and Operations Management all show good LCRs, indicating a strong interest among customers in these specializations.

# EDA – BIVARIATE ANALYSIS FOR NUMERICAL VARIABLES:




- There is a strong positive correlation between `'Total Visits'` and `'Page Views per Visit'`, indicating that customers who visit the website more frequently tend to view more pages per visit.

- Customers who spend `more time` on the `website` have a `higher LCR`, indicating that increasing the time spent on the website can lead to `higher conversion rates`.

# 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
  - 70:30 % ratio was chosen for the split.
- Feature scaling
  - Standardization method was used to scale the features
- Checking the correlations
  - Predictor variables which were highly correlated with each other were dropped (Lead_Origin_Lead Import and Lead Origin_Lead Add Form).

# MODEL BUILDING:

The steps to build the model are as follows:

- Creating Dummy Variables.
- Splitting the Dataset into train set and test set.
- Scaling of Features.
- Feature Elimination based on Correlation Check.
- Feature Selection Using RFE (Recursive Feature Elimination).
- Model building.
- Removing the less relevant variables based on RFE, p-values, and VIFs value.
- Evaluating the accuracy and other metrics of the model.

# FEATURE SELECTION

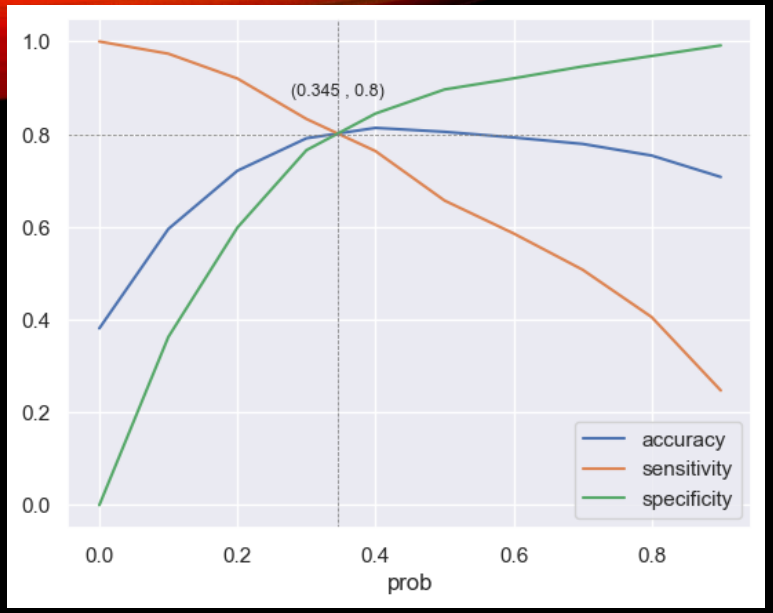
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome.
  - Pre RFE – 48 columns & Post RFE – 15 columns.
- 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, logm4 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

# MODEL EVALUATION:

*Tools used to evaluate the model are:*

1. Confusion Matrix
2. Accuracy
3. Sensitivity and Specificity
4. Threshold determination using ROC & Finding Optimal cutoff point.
5. Precision and Recall

# CONTD:

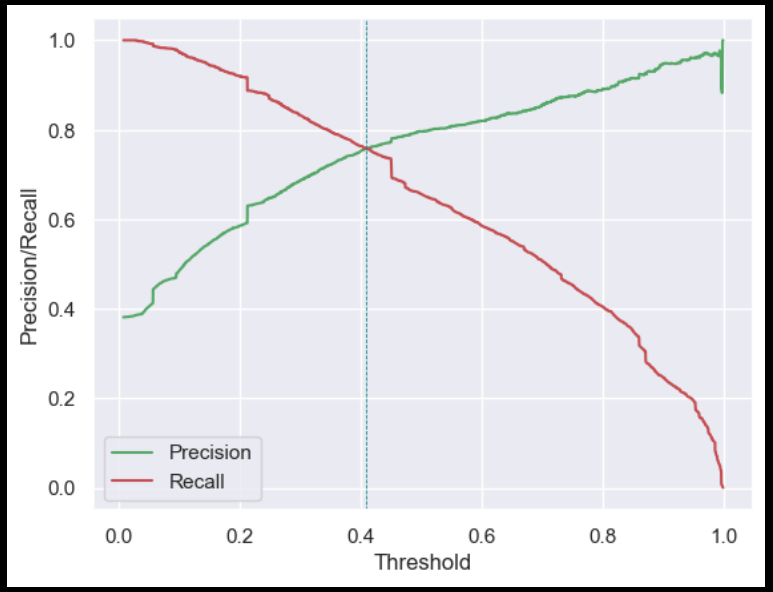



*Optimal Cut-Off Point:*

•Graph indicated the optimal cutoff point or probability.

•This analysis is necessary as it identifies the threshold that maintains a balance between sensitivity and specificity.

•Graph shows that point 0.345 (approx.) is the optimal cut-off value.

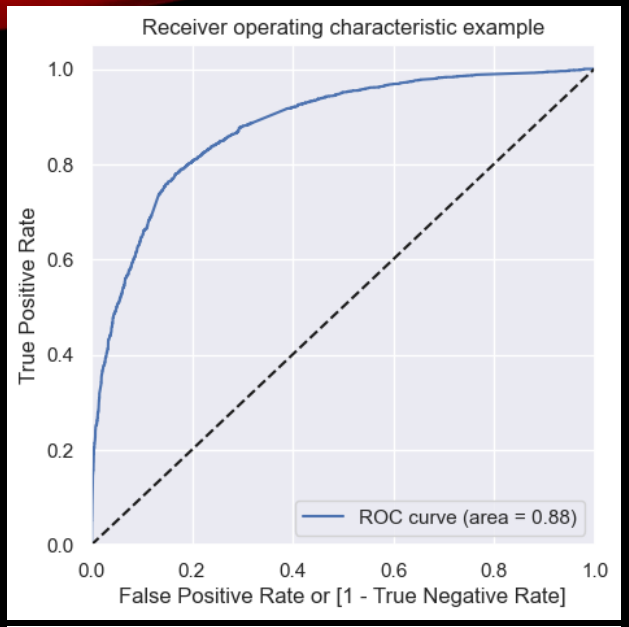


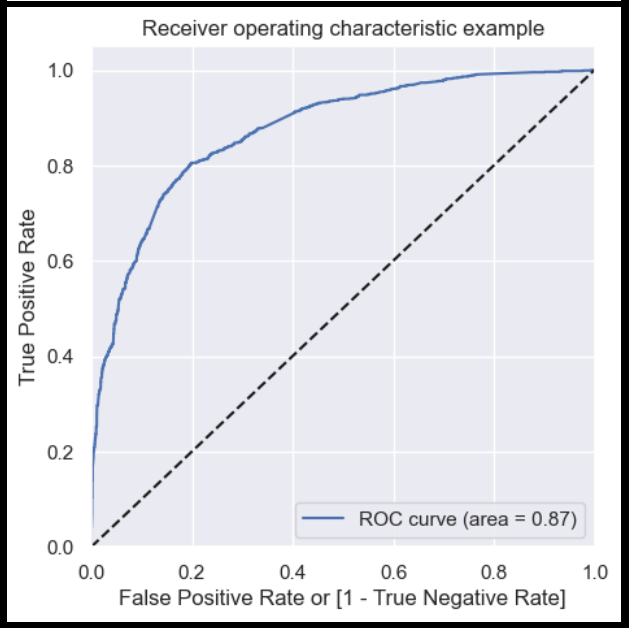

*Precision vs Recall Trade-Off:*

•Comparing the Precision-Recall view with the Specificity - Sensitivity view, we can conclude that the threshold value provides the best balance between these metrics.

•We can conclude that Precision v/s Recall curve achieves balance at an intersection that is the location where we find the optimal threshold value which in this case is 0.41 approx.

# MAKING PREDICTIONS ON TEST DATASET




*ROC Curve - Train Data Set:*

- *The Area under ROC curve was found to be 0 88 out of 1 indicating that the model is a good predictor.*
- *The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values.*

*ROC Curve - Test Data Set:*

- *The Area under ROC curve was found to be 0 87 out of 1 indicating that the model is a good predictor.*
- *The curve is plotted as close to the top left corner of the plot as possible, which indicates that the model has a high true positive rate and a low false positive rate at all threshold values*

# CONFUSION MATRIX & METRICS

```
# Predicted        not_converted  |  converted
# Actual                          |
# ----------------------------------------------
# not_converted       3230        |   772
# converted           492         |   1974
```

- For `Train Set`:
  - Accuracy: `0.8046`
  - Sensitivity: `0.8005`
  - Specificity: `0.8071`
  - False_positive_rate: `0.1929`
  - Precision: `0.7189`
  - Recall: `0.8005`
  - Negative_predictive_value: `0.8678`

```
# Predicted        not_converted  |  converted
# Actual                          |
# ----------------------------------------------
# not_converted       1353        |   324
# converted           221         |   874
```

- `For Test set`
  - Accuracy : `80.34%`
  - Sensitivity : 79.82% ≈ `80%`
  - Specificity : `80.68%`
  - false_positive_rate : `19.32%`
  - precision : `72.95%`
  - recall : `79.82%`
  - negative_predictive_value : `85.96%`

| | Prospect ID | Converted | Converted_Prob | final_predicted | Lead_Score |
|---|---|---|---|---|---|
| 0 | 4269 | 1 | 0.697934 | 1 | 70 |
| 1 | 2376 | 1 | 0.860665 | 1 | 86 |
| 2 | 7766 | 1 | 0.889241 | 1 | 89 |
| 3 | 9199 | 0 | 0.057065 | 0 | 6 |
| 4 | 4359 | 1 | 0.871510 | 1 | 87 |
| 5 | 9186 | 1 | 0.503859 | 1 | 50 |
| 6 | 1631 | 1 | 0.419681 | 1 | 42 |
| 7 | 8963 | 1 | 0.154531 | 0 | 15 |
| 8 | 8007 | 0 | 0.072344 | 0 | 7 |
| 9 | 5324 | 1 | 0.298849 | 0 | 30 |

- *Using a cut-off value of 0.345, the model achieved a sensitivity of 80.05% in the train set and 79.82% in 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 80.46%, which is in line with the study's objectives.*

# INSIGHTS:

***Based on the model evaluation:***

We know that the relationship between ln(odds) of 'y' and feature variable "X" is much more intuitive and easier to understand. The equation is:

ln(odds) = *1.0236 x const + 1.0498 x Total Time Spent on Website 1.259 x Lead Origin_Landing Page Submission + 0.9072 x Lead_Source_Olark Chat + 2.9253 x Lead Source_Reference + 5.3887 x Lead Source_Welingak Website + 0.9421 x Last Activity_Email_Opened - 0.5556 x Last Activity_Olark Chat Conversation + 1.2531 x Last Activity_Others + 2.0519 x Last Activity_SMS Sent 1.0944 x Specialization_Hospitality Management 1.2033 x Specialization_Others + 2.6697 x Current_Occupation_Working Professional*

***Train Data Set:***

- Accuracy: 80.57%
- Sensitivity: 79.72%
- Specificity: 81.08%

***Test Data Set:***

- Accuracy: 80.34%
- Sensitivity: 79.27%
- Specificity: 81.04%

# CONCLUSIONS:

- The final Logistic Regression Model has 12 features.
- Features that are contributing positively to predicting hot leads(leads that can be converted) in the model are:
  - Lead Source_Welingak Website
  - Lead Source_Reference -Current_occupation_Working Professional
- We noted  that the Optimal threshold/cutoff probability point was found to be 0.345.
- Converted probability predicted having a value greater than 0.345 will be predicted as Converted lead (i.e. Hot lead) while those smaller than 0.345 will be predicted as not Converted lead (Cold lead).
- Final model achieved an accuracy of 80.34% on the test data set, which is in line with the study's objectives as mentioned in the problem statement.

# RECOMMENDATIONS:

*In-Order to increase our Lead Conversion Rates following steps are to be taken:-*

- Focus on features such as 'Lead Source_Welingak Website', 'Current_Occupation_Working Professional', and 'Lead Source_Reference' with positive coefficients for targeted marketing strategies.
- Increasing the frequency of media usage such as Google ads or email campaigns can save time and increase the conversion rate.
- Leads whose Last Activity is SMS Sent or Email Opened tend to have a higher conversion rate and should be targeted more frequently.
- Referral leads generated by old customers have a significantly higher conversion rate and should be incentivized with discounts or other rewards to encourage more referrals.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

*Areas of improvement:-*

- Analyzing the behavior of customers who spend more time on the website can help improve the user experience and increase conversion rates, and company should focus on creating engaging content and user-friendly navigation to encourage customers to spend more time on the website.
- Understanding the most popular specializations can help tailor course offerings and marketing campaigns to specific groups of customers. Providing targeted content and resources for popular specializations such as Marketing Management and HR Management can also help attract and retain customers in those fields

# RECOMMENDATIONS:

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:

- Lead Source_Welingak Website: 5.39
- Lead Source_Reference: 2.93
- Current_occupation_Working Professional: 2.67
- Last Activity_SMS Sent: 2.05
- Last Activity_Others: 1.25
- Total Time Spent on Website: 1.05
- Last Activity_Email Opened: 0.94
- Lead Source_Olark Chat: 0.91

We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:

- Specialization in Hospitality Management: -1.09
- Specialization in Others: -1.20
- Lead Origin of Landing Page Submission: -1.26

# THANK YOU