Lead Scoring Case Study

by,

Jobish Jose | Karthik Ragav Kumaresan | Joel Anthony Jose

Problem Statement & Business Understanding

- X Education is an online education company that sells courses to industry professionals.
- The company gets leads from its website, search engines, and past referrals.
- The typical lead conversion rate at X Education is around 30%. The CEO wants to improve the lead conversion rate to 80%.
- The company requires a model 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.

Business Objectives

- The goal of this project is to help X Education improve its lead conversion rate and increase sales.
- By identifying the most potential leads, the sales team can focus their efforts on communicating with the leads that are most likely to convert. This will lead to a more efficient and effective sales process.
- 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 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.
- Analysis and Model Building is done using Python on a Jupyter notebook.

Understanding Dataset

- A dataset with more than 9000 data points are provided(9240 rows * 37 columns).
- The dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, 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.

Approach

- Data Reading and Understanding
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Data Preparation
- Model Building
- Model Evaluation
- Predictions on Test Data
- Recommendations

Data Reading & Understanding

- The dataset contains 9240 rows and 37 columns.
- Checked the data types of all the columns.
- Gone through the statistical summary of all the numerical columns.
- Checked for null values and duplicates – There are a few columns with high number of null/ missing values and no duplicate rows present in the dataset.

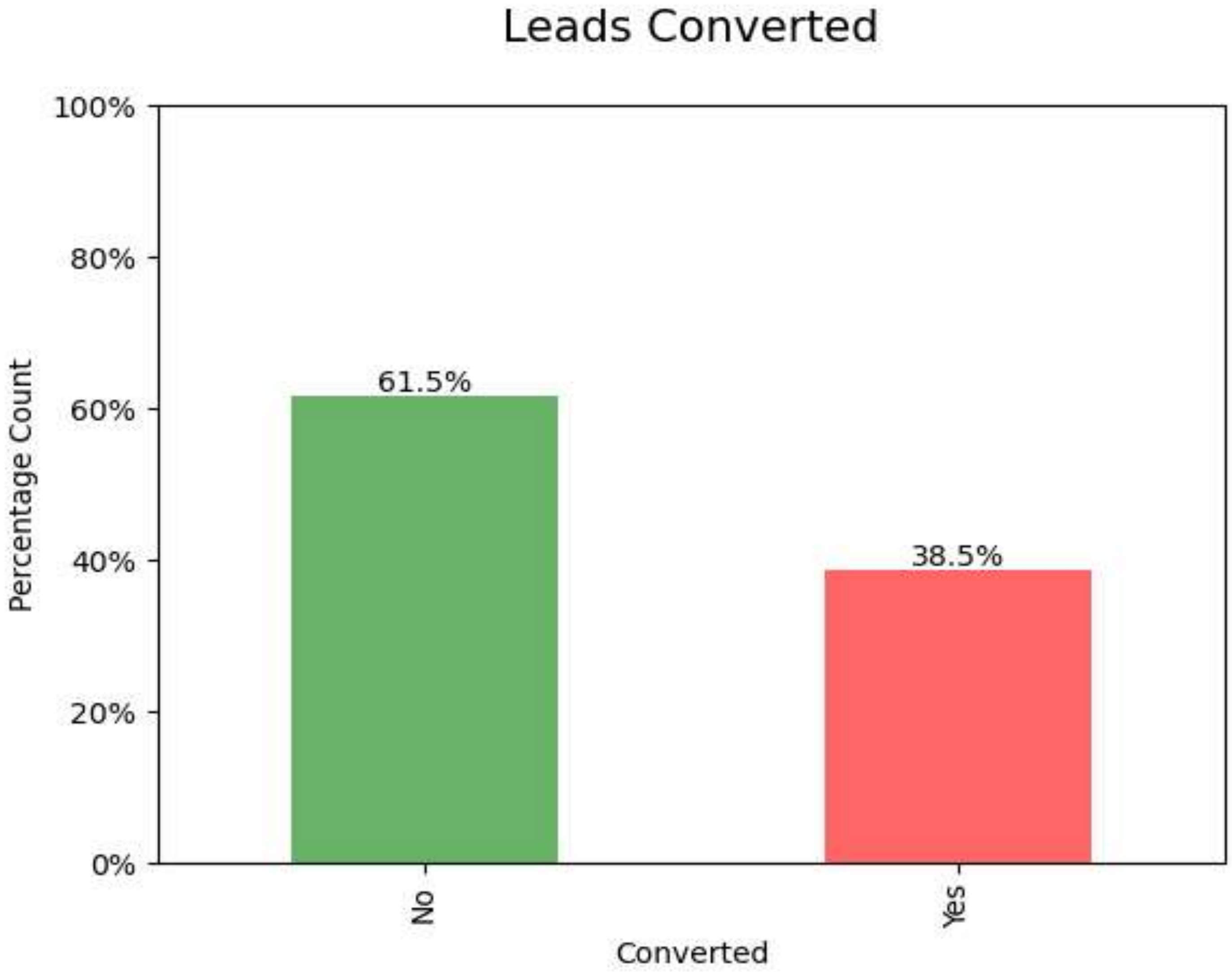
Data Cleaning

- The "Select" level in certain categorical variables was replaced with 'NaN' to signify null values, as it denoted instances where customers made no selections.
- Columns with over 40% missing values were removed, along with 'Prospect ID' and 'Lead Number' due to unique values in each row.
- Additionally, columns with constant values across all rows were also dropped.
- Missing values in categorical columns were addressed by either imputation or column removal, depending on category distribution and data balance. - Imputation was used if the categories are evenly distributed and dropped the columns if imputation would result in skewed data.
- Numerical columns with missing values were imputed using the mode after assessing their distribution.

Data Cleaning

- Columns which do not add any usable information(‘Last Notable Activity”) have been dropped.
- Columns with highly skewed data have been also dropped to prevent from biased estimates.
- Outliers in ‘TotalVisits’ and ‘Page Views Per Visit’ were treated by capping and flooring.
- Standardized the cases of values(‘Google’ and ‘google’) in columns where required.(‘Lead Source’)
- Low frequency values were grouped together to a new category called “Others’.
- Binary categorical variables were mapped accordingly.
- Columns with lengthy names were renamed to shorter names for better readability and maintainability.

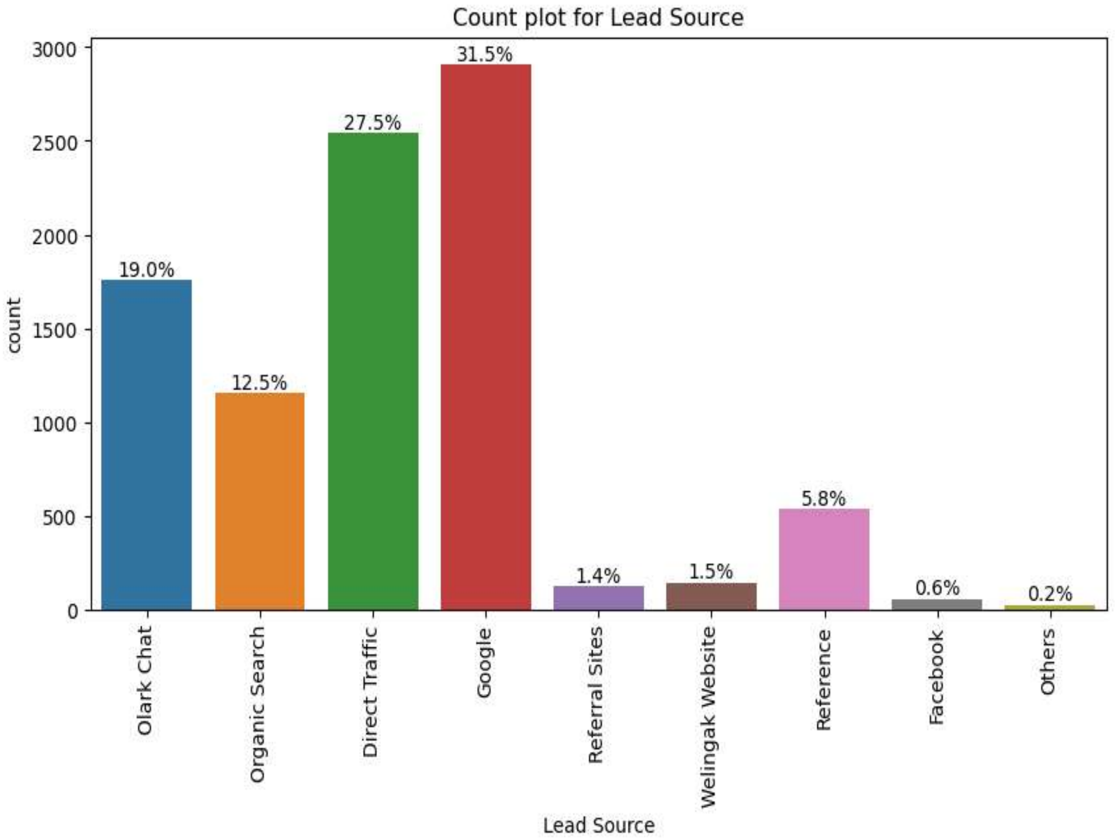
Exploratory Data Analysis(EDA)

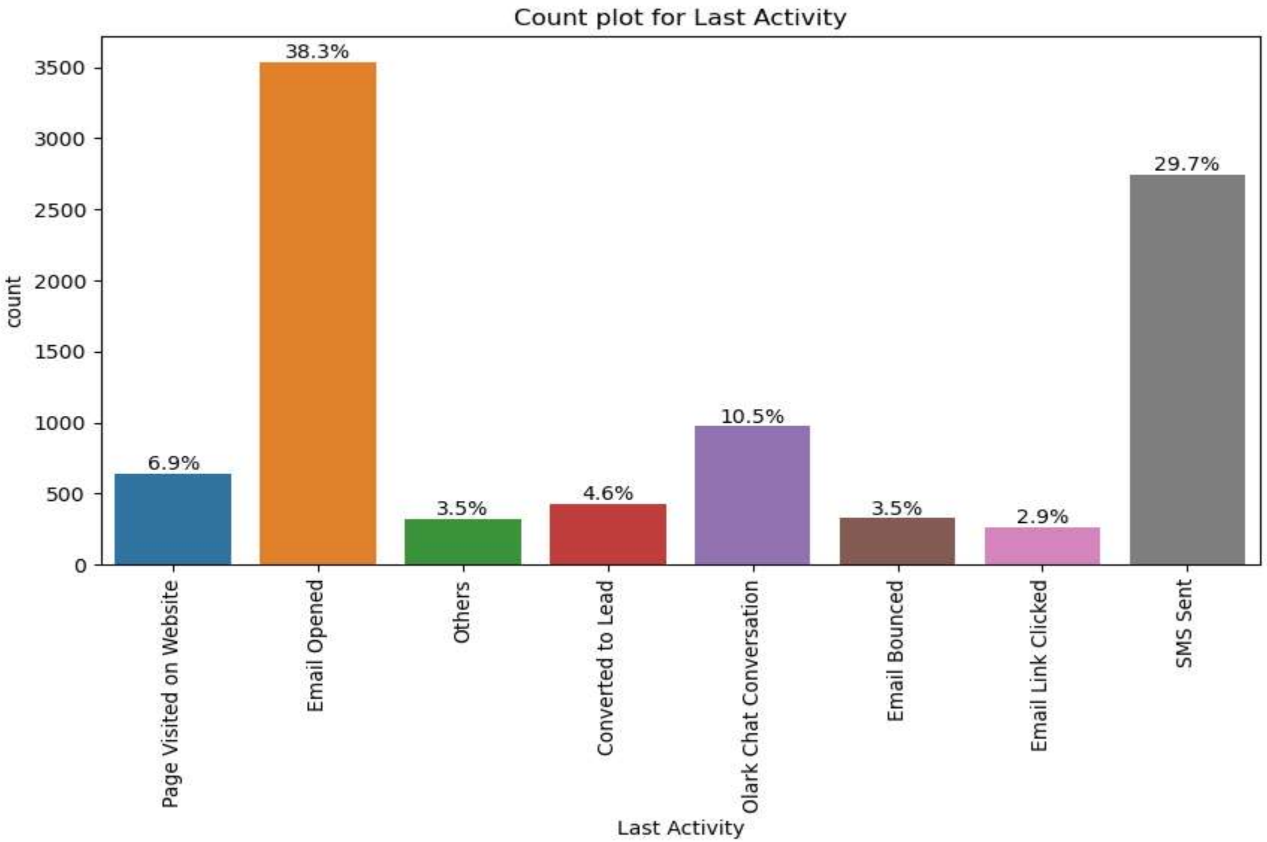


Leads Conversion Rate (LCR) is 38.5%.

Exploratory Data Analysis(EDA)

Univariate Analysis – Categorical Variables



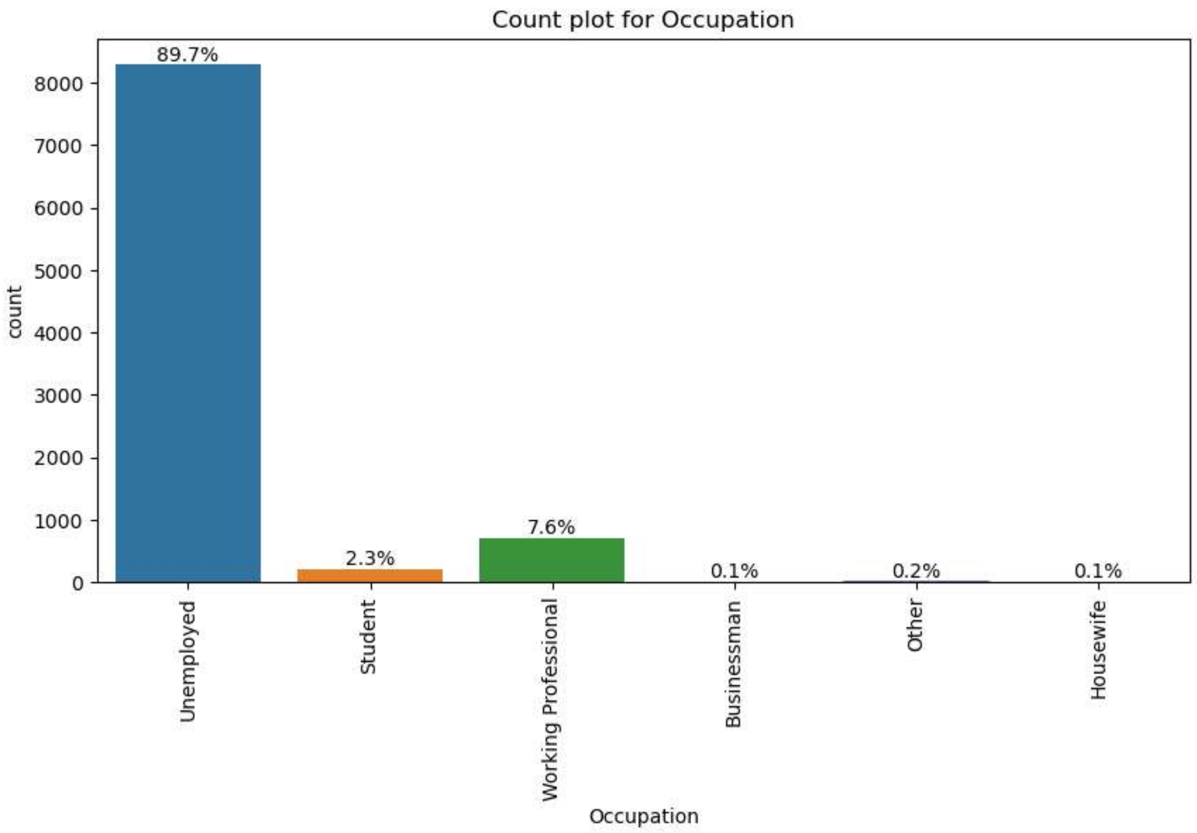


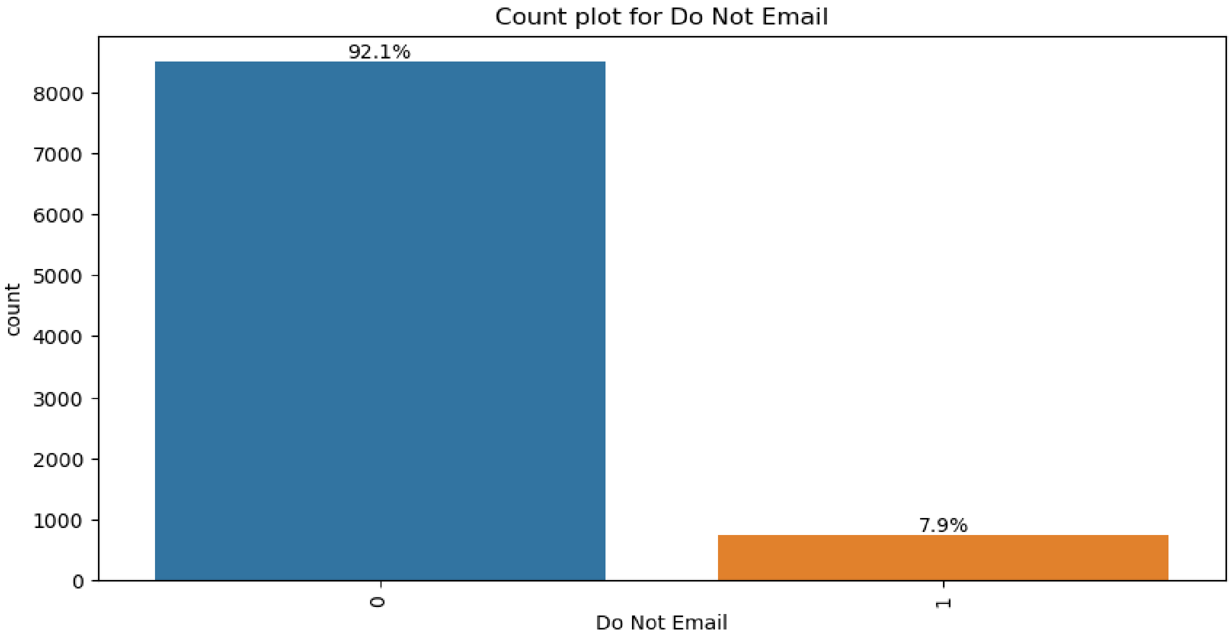
Lead Source - 'Google' leads with 31.5% customers, closely followed by 'Direct Traffic'(27.5%) and 'Olark Chat'(19%)

Last Activity - 'Email Opened' leads the list with 38.3% followed by 'SMS Sent'(29.7%)

Exploratory Data Analysis(EDA)

Univariate Analysis – Categorical Variables



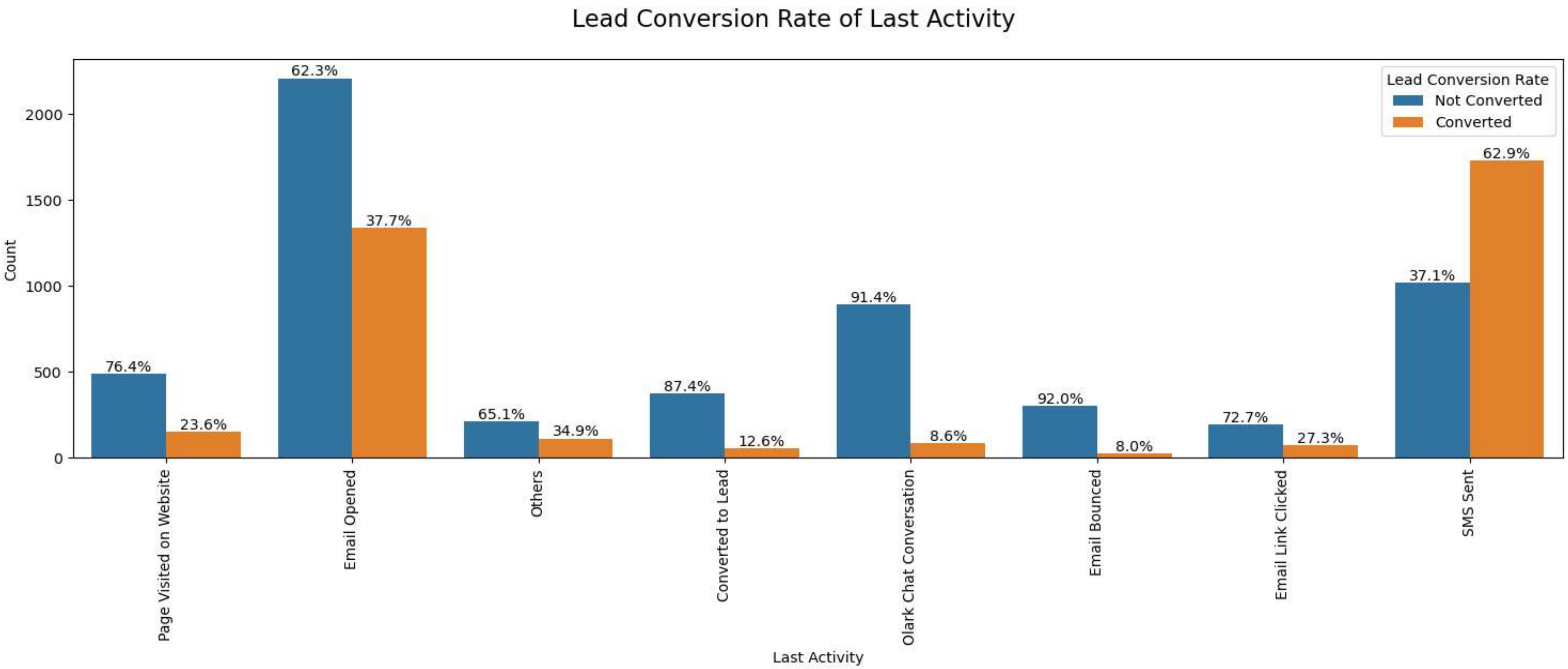


Occupation - 89.7% of the customers are 'Unemployed', while 7.6% are 'Working Professional' and 2.3% are 'Student'.

Do Not Email - 7.9% of the customers opted out from Emails; They do not want to get notified about the course through email.

Exploratory Data Analysis(EDA)

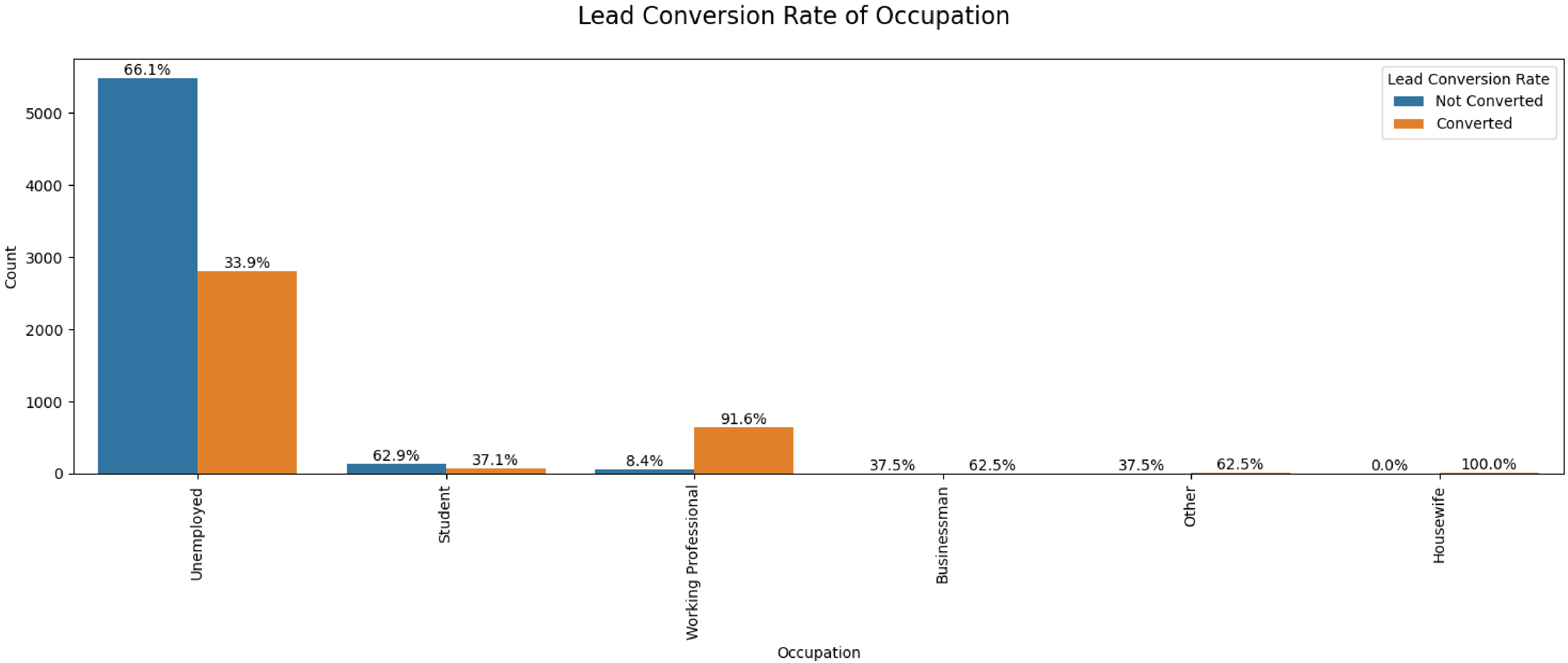
Bivariate Analysis – Categorical Variables



Last Activity - The most effective category is 'SMS Sent' with a lead conversion rate of 62.9%, followed by 'Email Opened'(37.7%).

Exploratory Data Analysis(EDA)

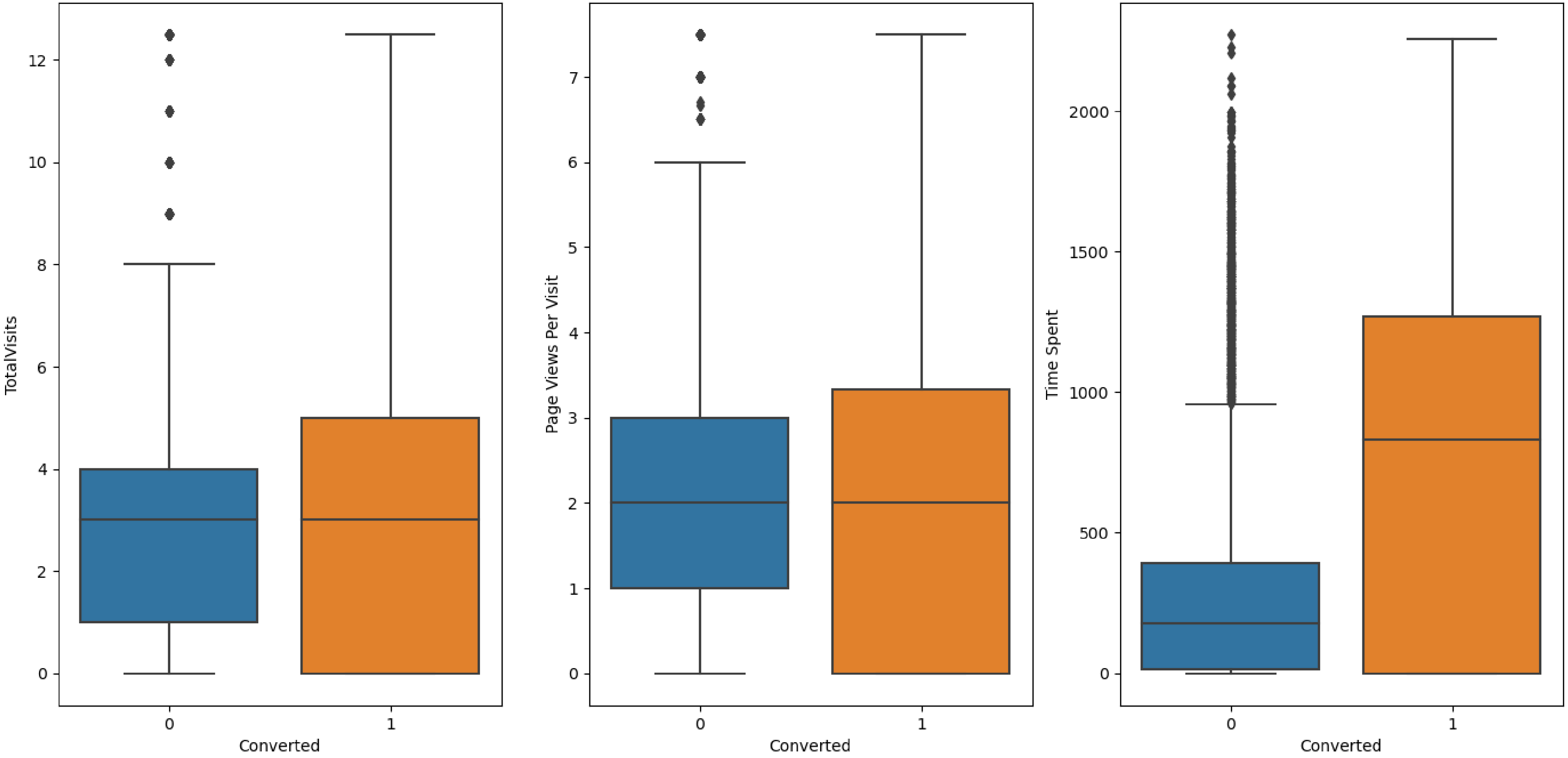
Bivariate Analysis – Categorical Variables



Occupation - 'Working Professional' has 91.6 % lead conversion rate. Another effective category is 'Unemployed' with 33.9% of lead conversion rate.

Exploratory Data Analysis(EDA)

Bivariate Analysis – Numerical Variables



Time Spent - Customers who spend more time on website have a higher lead conversion rate.

Data Preparation

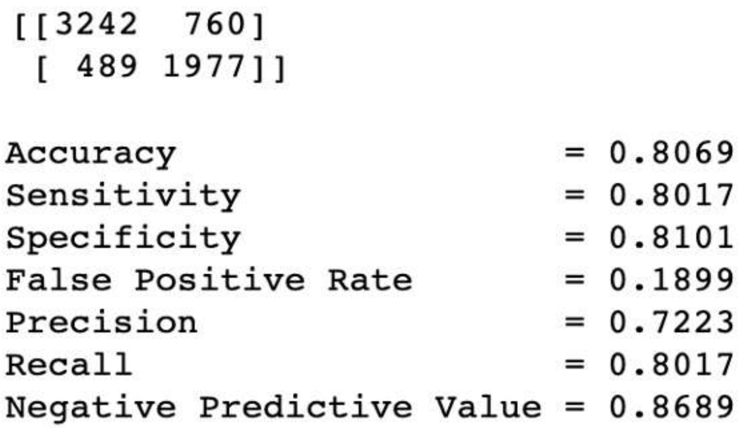
- Dummy variables were created for the categorical variables. ('Lead Origin', 'Lead Source', 'Last Activity', 'Specialization’ and 'Occupation’).
- Split the dataset into Train and Test sets. (70 : 30 ratio)
- Feature scaling was performed using Standard Scaler.
- Plotted a heatmap to check the correlations between the predictor variables and dropped the variables which were highly correlated with each other. ('Lead Origin_Lead Import' and 'Lead Origin_Lead Add Form’)

Model Building

- Used Recursive Feature Elimination(RFE) to pick the most important variables and thus reduced number of variables from 45 to 15.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Two models were built before reaching the final model – Model 3(lrm3).
- Model 3 (lrm3) looked stable with p-values of all variables less than 0.05 indicating that the variables are statistically significant predictors, and VIF all the variables less than 5 indicating that there are no high multicollinearity among the variables.
- Model 3 has been selected as the final Logistic Regression Model for making further predictions.

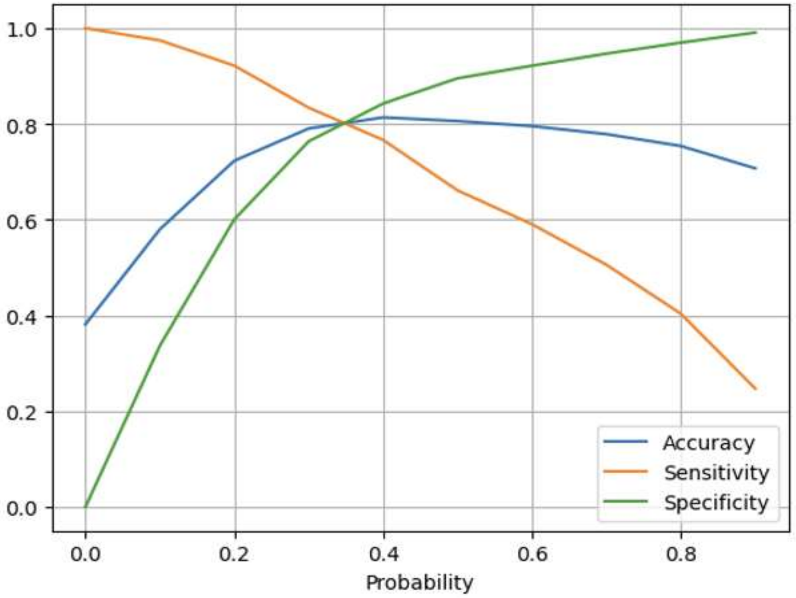
Model Evaluation

Confusion Matrix & Evaluation Metrics with 0.35 as cut – off.

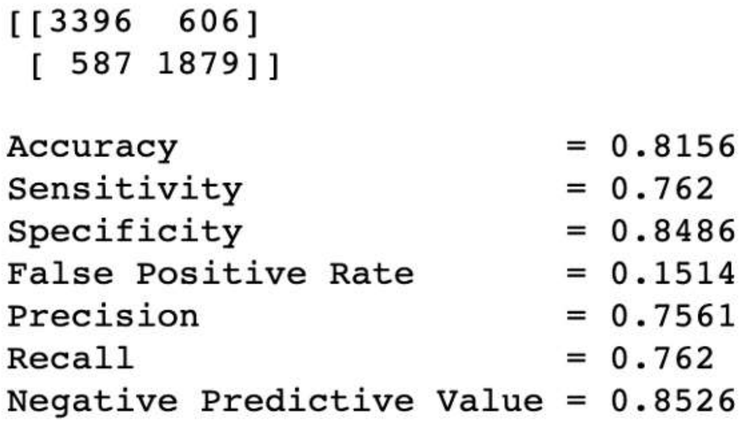


```
[[3242  760]
 [ 489 1977]]

Accuracy                  = 0.8069
Sensitivity               = 0.8017
Specificity               = 0.8101
False Positive Rate       = 0.1899
Precision                 = 0.7223
Recall                    = 0.8017
Negative Predictive Value = 0.8689
```

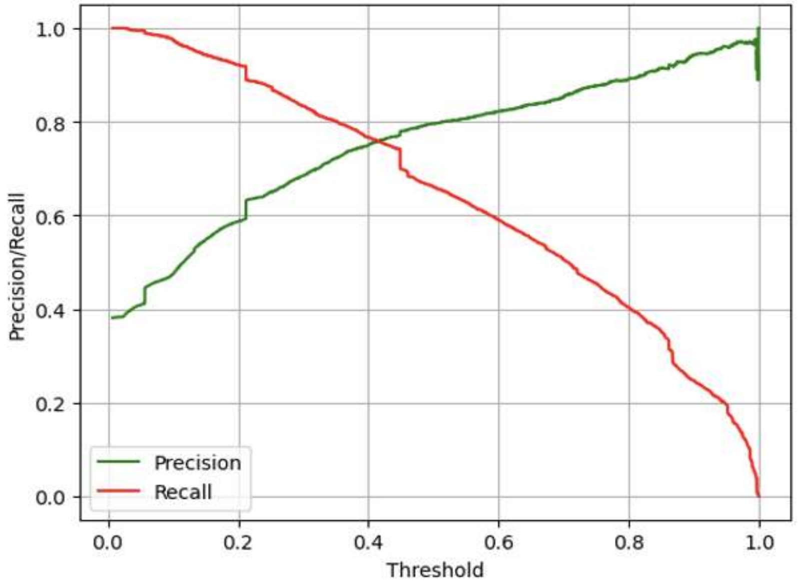


Confusion Matrix & Evaluation Metrics with 0.41 as cut – off.



```
[[3396  606]
 [ 587 1879]]

Accuracy                  = 0.8156
Sensitivity               = 0.762
Specificity               = 0.8486
False Positive Rate       = 0.1514
Precision                 = 0.7561
Recall                    = 0.762
Negative Predictive Value = 0.8526
```

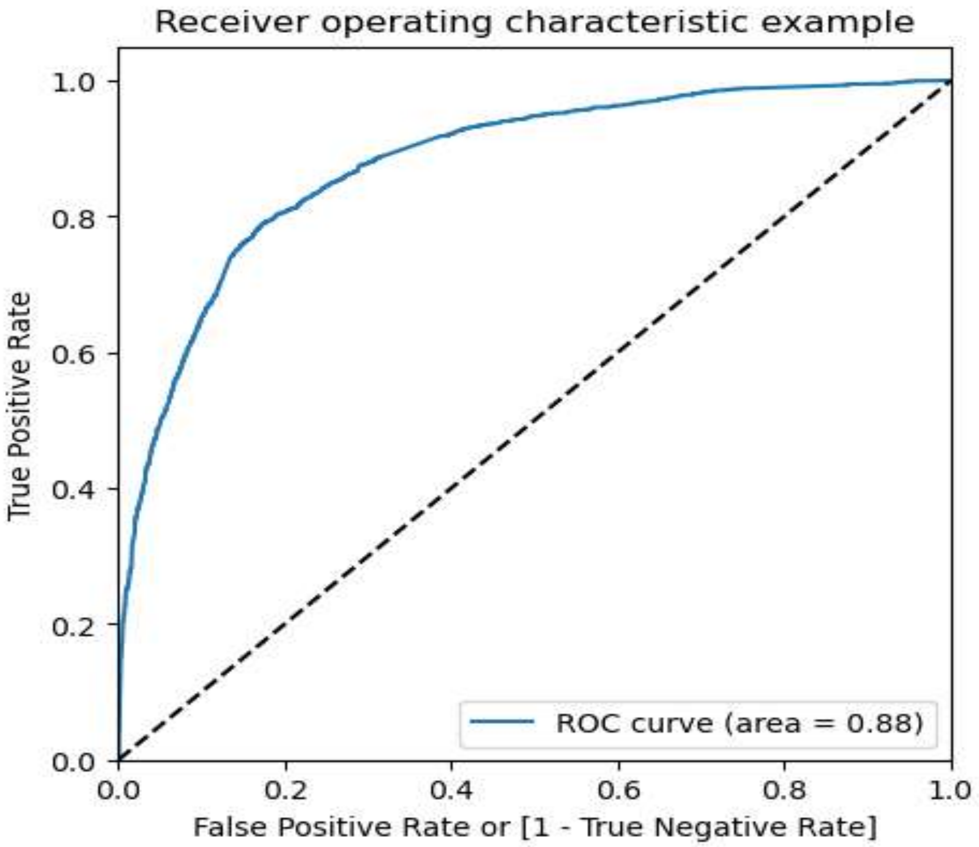


- Accuracy Sensitivity Specificity cut-off is 0.35.
- Precision Recall cut-off is 0.41.

- **Following a comparison between the Confusion Matrix and Evaluation Metrics, the decision was made to proceed with a 0.35 cut-off threshold as it yielded superior results.**

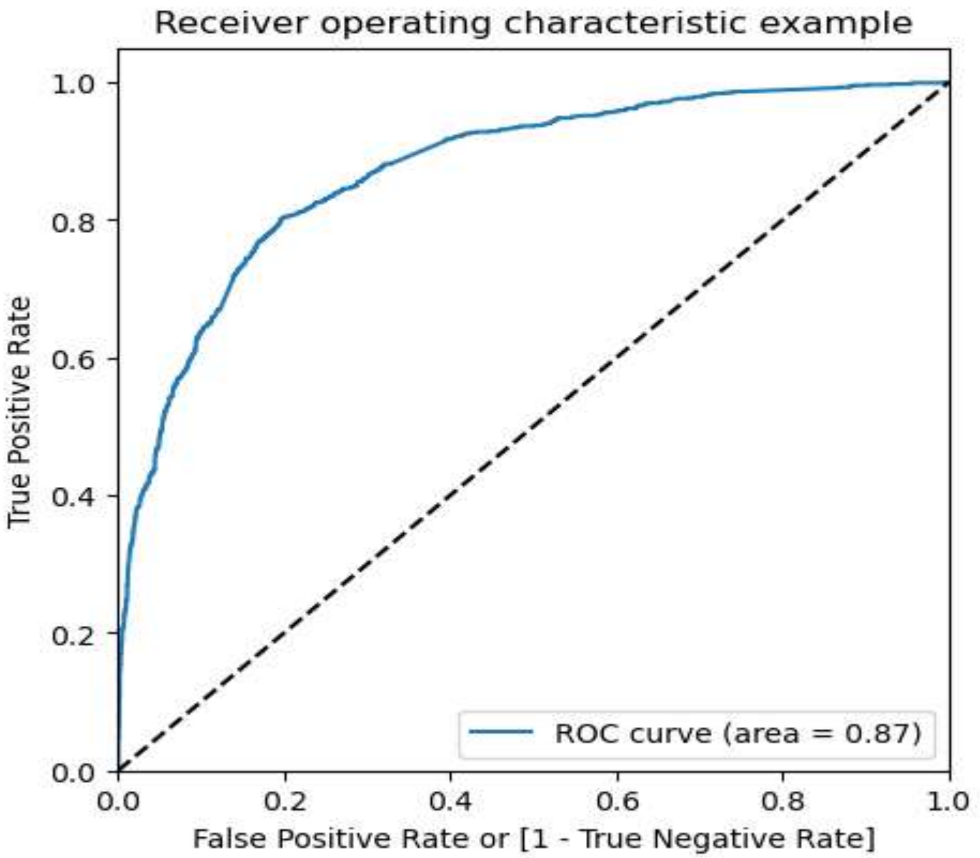
Model Evaluation

ROC Curve – Train Data Set



With an area under the curve of 0.88, it suggests that the model is doing a commendable job in effectively distinguishing between classes.

ROC Curve – Test Data Set

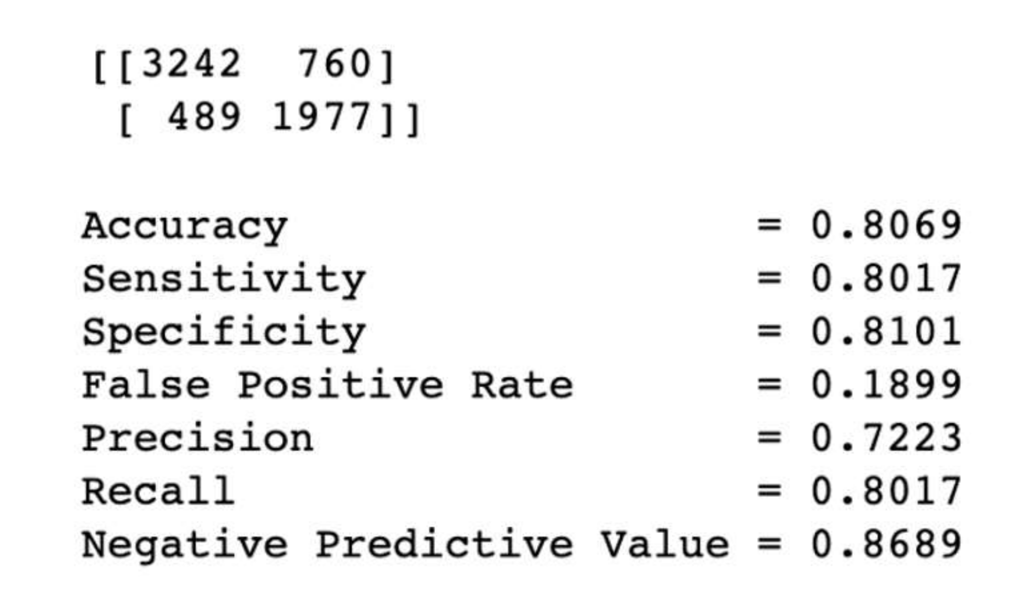


With an area under the curve of 0.87, it suggests that the model is doing a commendable job in effectively distinguishing between classes.

Model Evaluation

Confusion Matrix and Evaluation Metrics

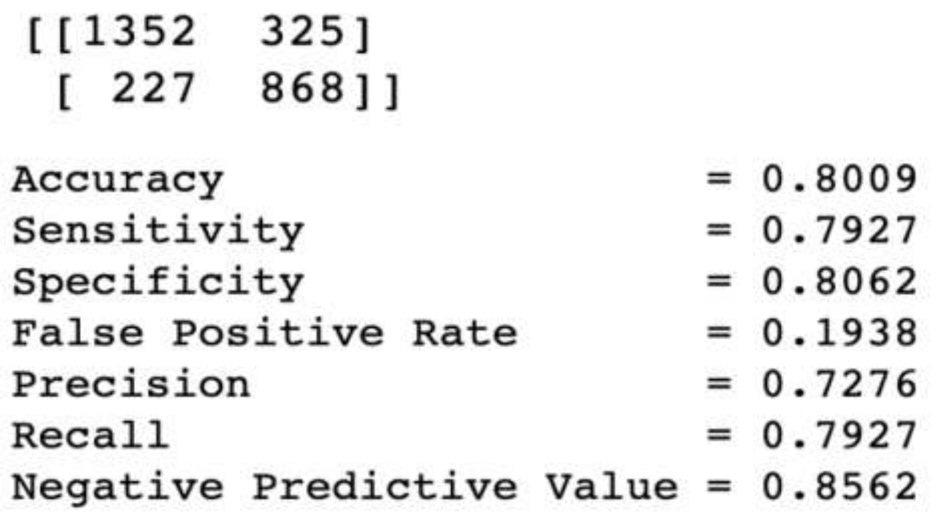
Train Data Set



```
[[3242  760]
 [ 489 1977]]

Accuracy                  = 0.8069
Sensitivity               = 0.8017
Specificity               = 0.8101
False Positive Rate       = 0.1899
Precision                 = 0.7223
Recall                    = 0.8017
Negative Predictive Value = 0.8689
```

Test Data Set



```
[[1352  325]
 [ 227  868]]

Accuracy                  = 0.8009
Sensitivity               = 0.7927
Specificity               = 0.8062
False Positive Rate       = 0.1938
Precision                 = 0.7276
Recall                    = 0.7927
Negative Predictive Value = 0.8562
```

The evaluation metrics in both train and test sets are really close to each other which indicates the model is performing consistently. This consistency serves as an indicator of the model's reliability, indicating that it does not suffer from overfitting and responding really good to new data which is a very important aspect while considering the deployment of the model in real-world applications. These findings reflect positively on the model's performance, instilling confidence in its capacity to make precise predictions in future scenarios.

Insights

- Leads generated from the Welingkak Website are more likely to convert than leads generated from other sources. The coefficient for Lead Source_Welingak Website is the highest, indicating that this is the most important factor in predicting whether a lead will convert.
- Leads who come from a referral are also more likely to convert than leads who come from other sources. The coefficient for Lead Source_Reference is also positive, though smaller than the coefficient for Lead Source_Welingak Website.
- Leads who are working professionals are more likely to convert than leads who are not working professionals. The coefficient for Occupation_Working Professional is positive, suggesting that this is another important factor in predicting whether a lead will convert.
- Leads who have recently received an SMS are more likely to convert than leads who have not received an SMS. The coefficient for Last Activity_SMS Sent is positive, suggesting that this is another factor that can increase the chances of a lead converting.

Insights

- Leads who have recently opened an email are more likely to convert than leads who have not opened an email. The coefficient for Last Activity Email Opened is also positive, suggesting that this is another factor that can increase the chances of a lead converting.
- Leads who spend more time on the website are more likely to convert.
- Leads who click on a link in an email are more likely to convert.
- Leads who chat with a representative from Olark chat are more likely to convert.
- Leads who visit the website multiple times are more likely to convert.
- The 'Specialization_Hospitality Management' and 'Specialization_Others' features have negative coefficients indicating that leads with these specializations are less likely to achieve the desired outcome. These specializations may be associated with lower conversion rates.
- Leads who come through a 'Landing Page Submission' are less likely to convert

Recommendations

- Focus on leads from the Welingkar Website - Given that leads from the 'Welingkar Website' have the highest likelihood of conversion, allocate a significant portion of marketing budget and efforts to drive more traffic to the website. Optimize the website to capture and nurture leads effectively.
- Strengthen Referral Programs - Leads from referrals are also more likely to convert; Encourage and incentivize existing customers and partners to refer potential leads to the business. Implement a referral program with incentives to motivate existing customers to refer more frequently.
- Target working professionals - Working professionals are more likely to convert, so marketing and sales efforts should be tailored to this audience. Develop specialized campaigns or content tailored to their needs and interests. Highlight how the products or services can enhance their careers.
- Send SMS and email reminders - Sending SMS and email reminders can help to keep leads engaged and increase the chances of them converting. Make sure to send these reminders at the right time and with the right content.
- Enhance Website Experience - Given that leads who spend more time on the website are more likely to convert, encourage leads to spend more time on the website by offering informative, engaging, and relevant content. Implement chatbots or live chat support to assist visitors and guide them towards conversion actions.

THANK YOU