Lead Scoring Case Study

Presented By:

Nipun Garg
Monika Sekar
Abin Mathai

Contents

Problem Statement

X Education sells online courses to industry professionals. 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. Moreover, the company also gets leads through past referrals.

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%. Although X Education gets a lot of leads, its lead conversion rate is very poor.

Business Objective

X Education needs help to identify the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company requires a model wherein you 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%.

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.

Solution Approach

- Data Sourcing, Cleaning and Visualising
  - Reading and understanding data
  - Cleaning null values and dropping imbalance columns
  - Univariate and Bivariate analysis
- Data Preparation, Feature Scaling and Train-Test split
  - Creating dummy variables for categorical columns
  - Splitting data into train and test set
  - Feature scaling of numerical columns
- Model Building and Model Evaluation
  - Feature selection using RFE
  - Dropping features with high p-value and vif
  - Model evaluation using sensitivity-specificity and precision and recall
  - Making prediction on test set
- Conclusion
  - Determining the features which contribute most to a lead getting converted.
  - Determining lead scores and identifying the hot leads having scores above 70.

Reading Dataset

- This 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.
- Total number of rows = 9240, Total number of columns = 37

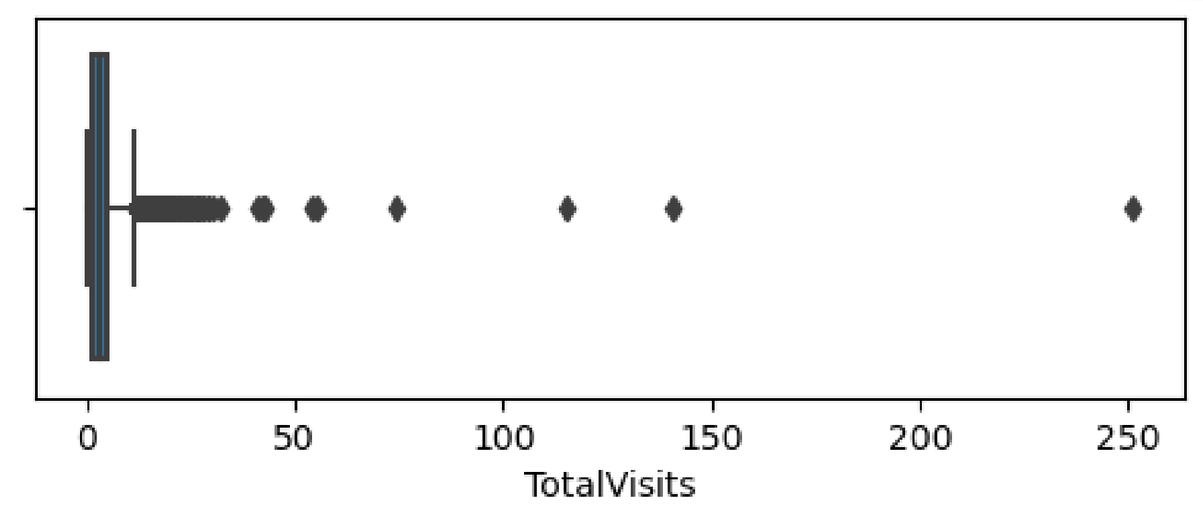
Data Cleaning

- There are 7 columns in 'leads.csv' with more than 40% null values in them. These columns were dropped as they could create problems in analysis.
- After checking value counts in 'City', 'Country', 'Tags', and 'What matters most to you in choosing a course' columns, we found that there is a high number of missing values and data imbalance in these columns. So these columns were dropped.
- Null values in 'Specialization' and 'What is your current occupation' is imputed with 'Not Provided'.
- Combined similar lead sources in one category to prevent unnecessary dummy columns.
- Dropped the null values rows in the columns 'TotalVisits' and 'Lead Source'.
- There were 12 columns with a single value or high data imbalance. These columns were dropped as they were not useful in the analysis.

Outliers Analysis

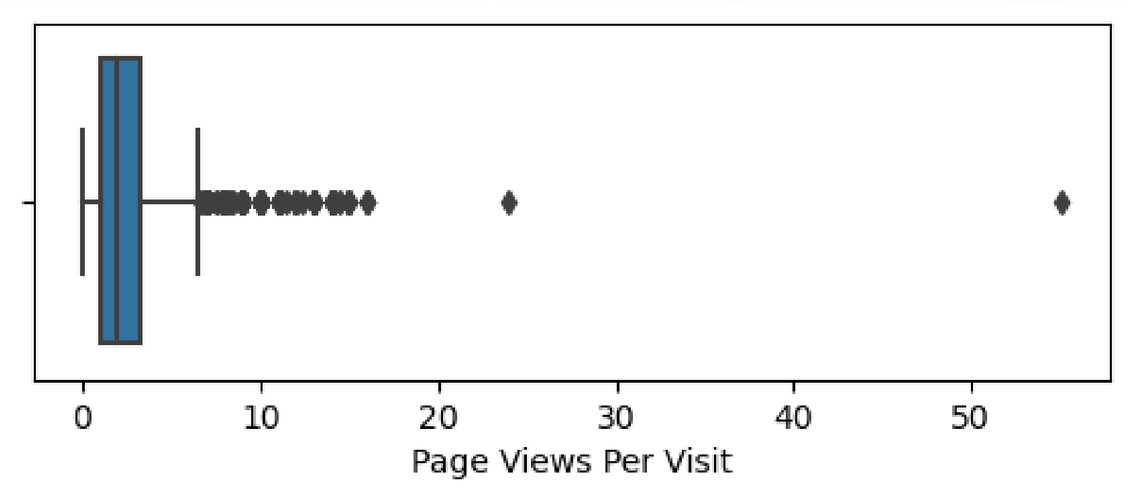
- **TotalVisits**

Outliers are present in 'TotalVisits'. Outliers are capped at the 99th percentile using soft capping.



- **Pages Views Per Visit**

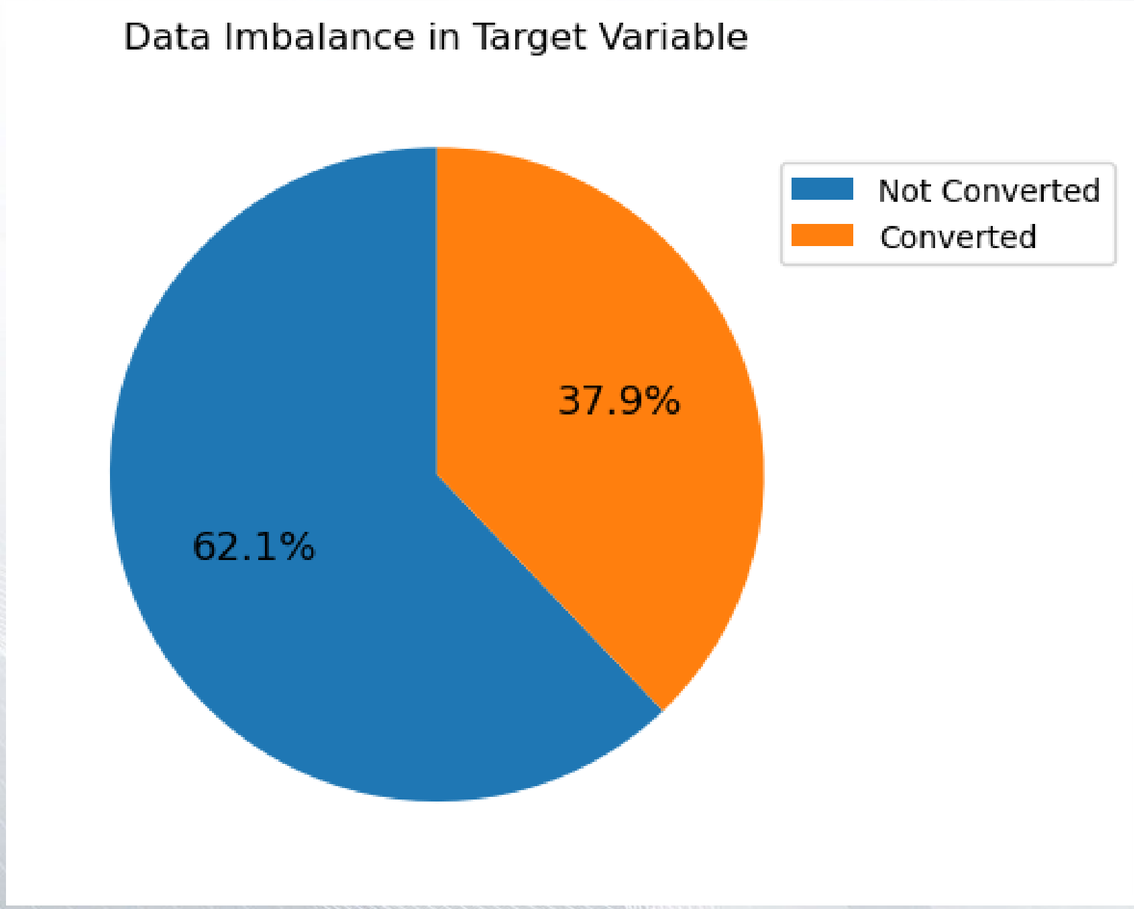
Outliers are present in 'Pages Views Per Visit'. Outliers are capped at the 99th percentile using soft capping.



Univariate Analysis

Data Imbalance in Converted (Target) Variable

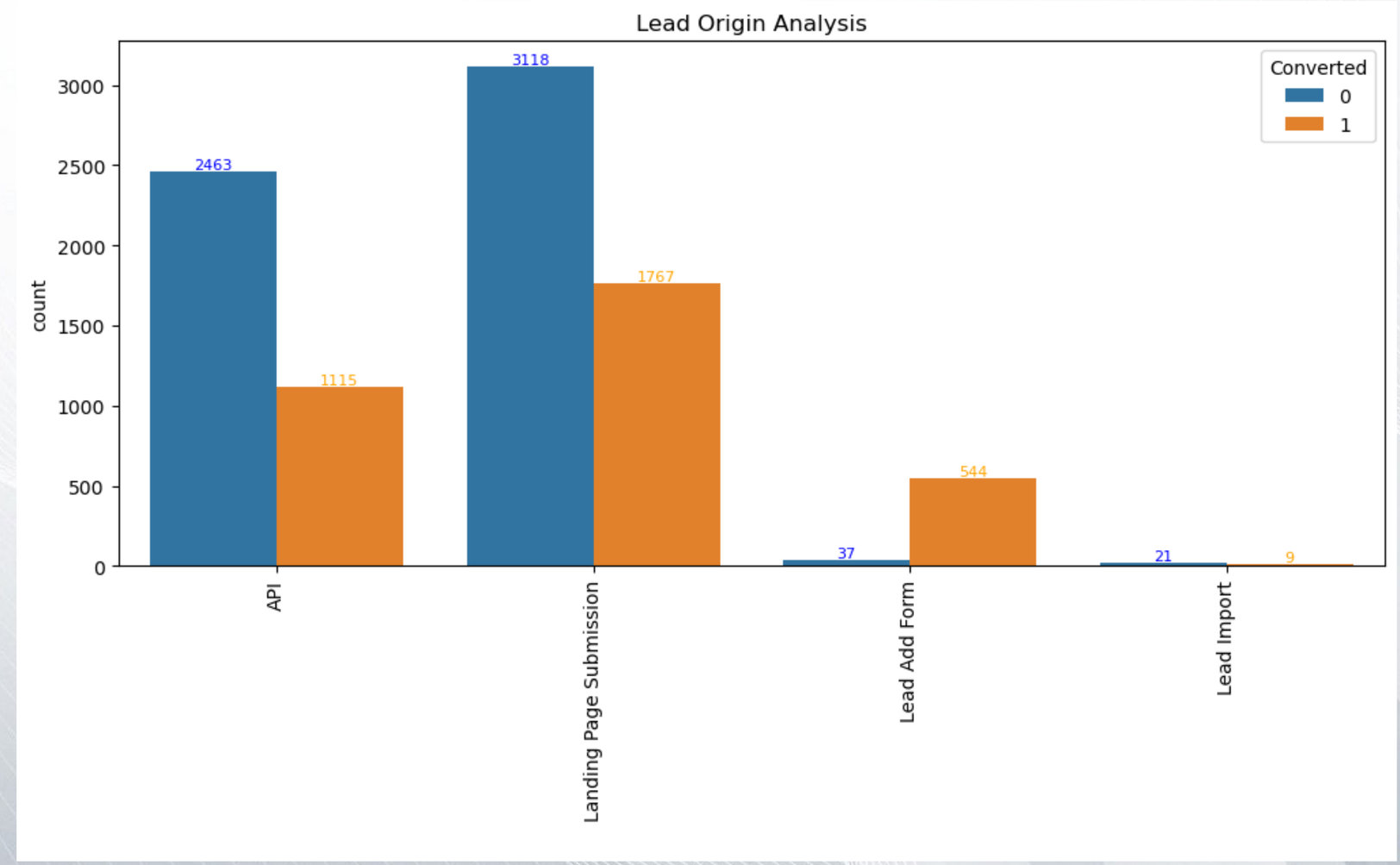
- We can see there is some data imbalance in the Converted column.



Univariate Analysis

Lead Origin Analysis

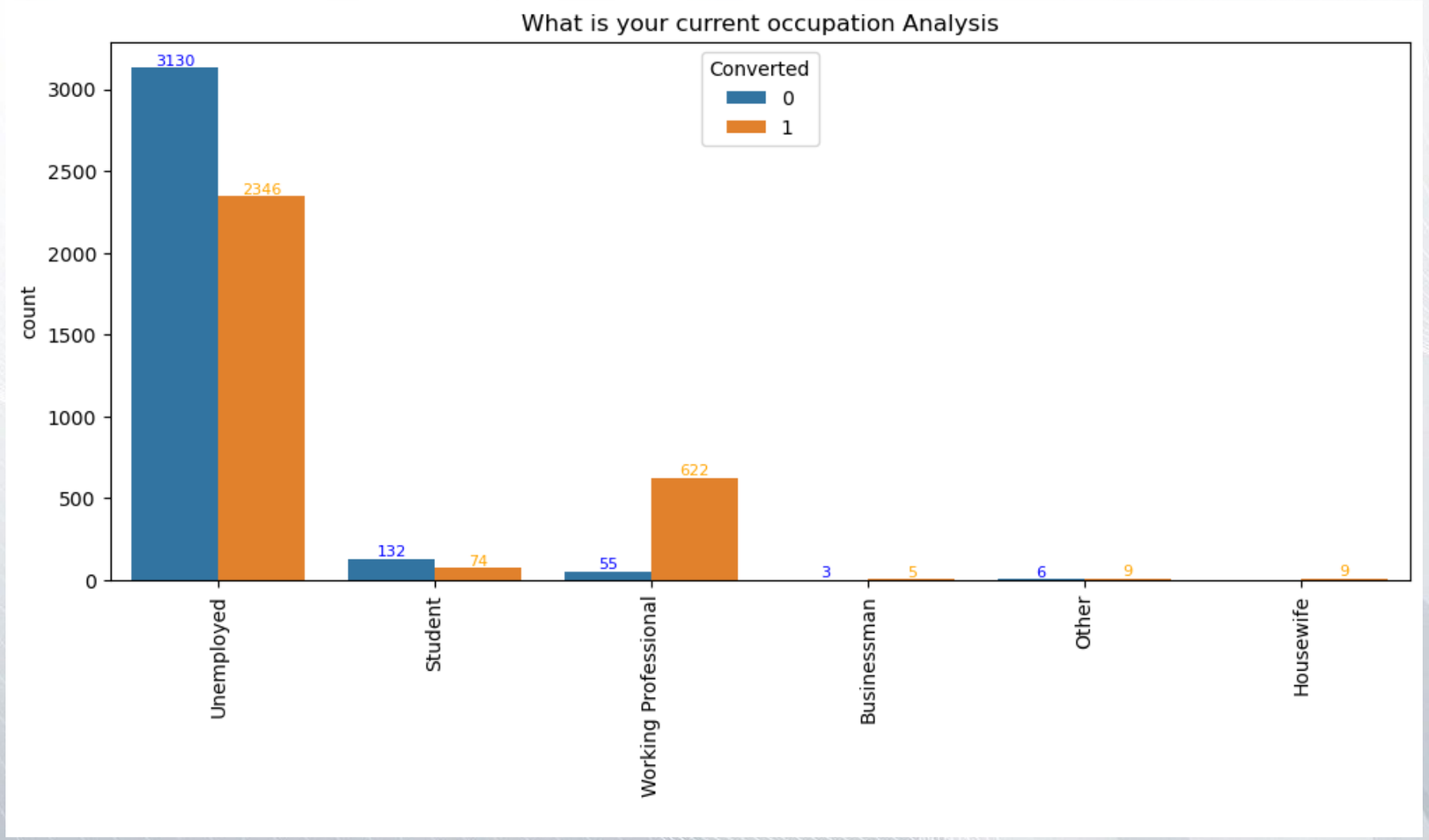
- Most of the leads identified were from Landing Page Submission.
- Customers from the Lead Add Form have the highest conversion rate.



Univariate Analysis

What is your current occupation Analysis

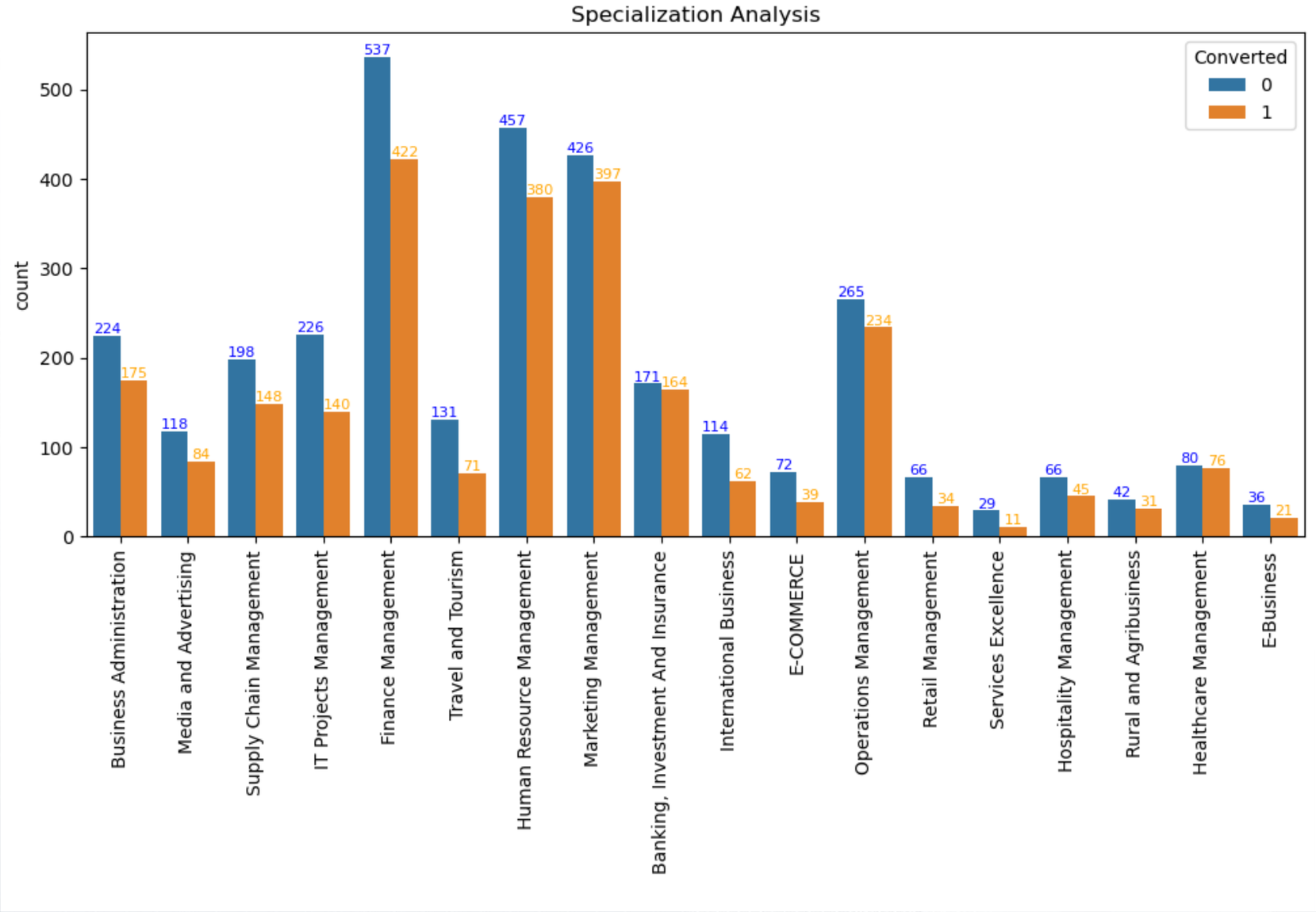
- Most of the leads are from Unemployed customers.
- Leads from Working Professionals have an excellent conversion rate of 91.8%.



Univariate Analysis

Specialization Analysis

- Management courses are more in demand than other courses.
- Human Resource Management, Marketing Management, Banking, Investment And Insurance, Operations Management, and Healthcare Management courses have higher conversion rates.

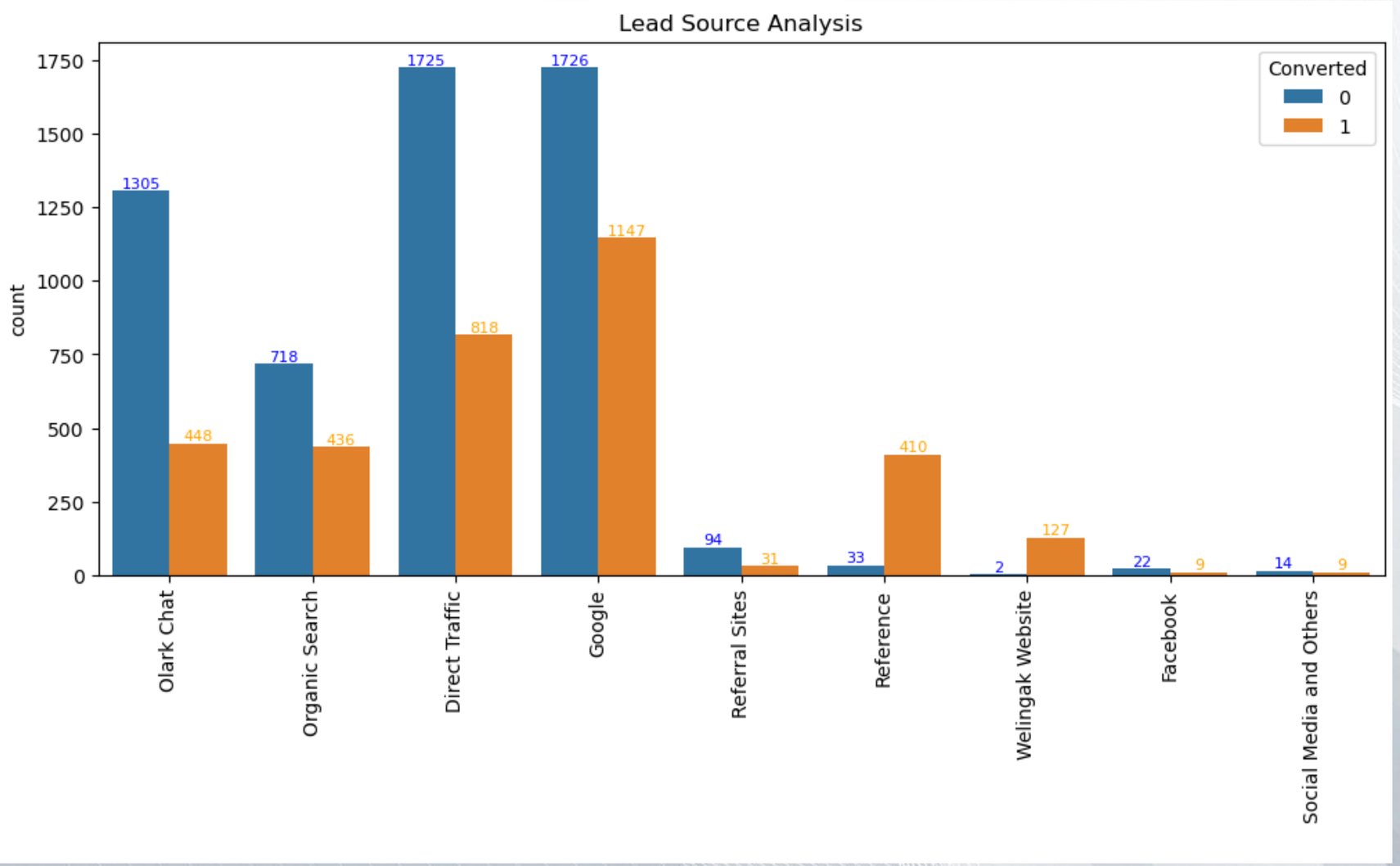


Univariate Analysis

Lead Source Analysis

- Most of the leads are coming from Google and Direct Traffic

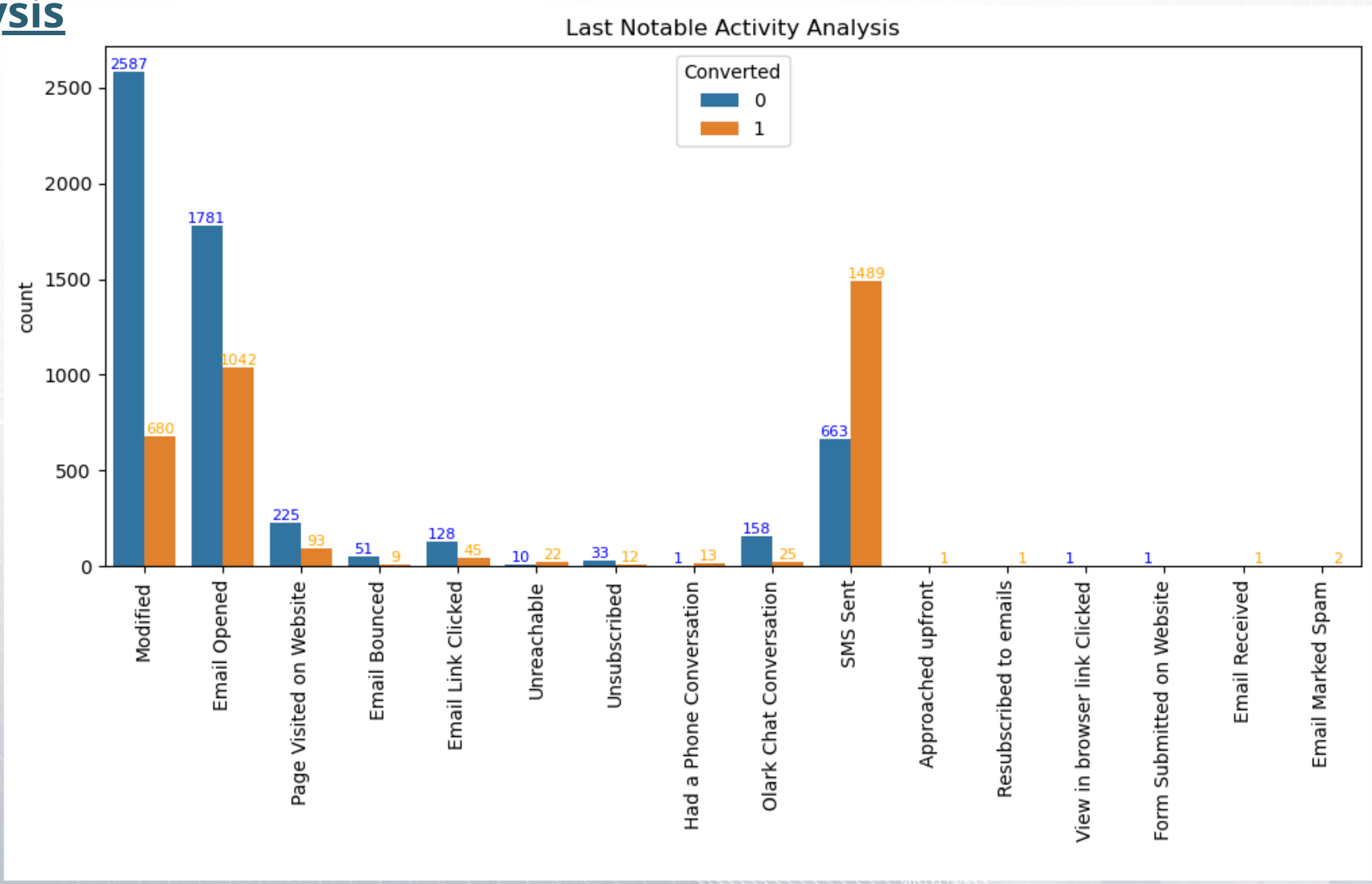
- Leads from Reference and Welingak Website have the highest conversion rate. X Education should run more referral programs to get more leads from existing students.



Univariate Analysis

Last Notable Activity Analysis

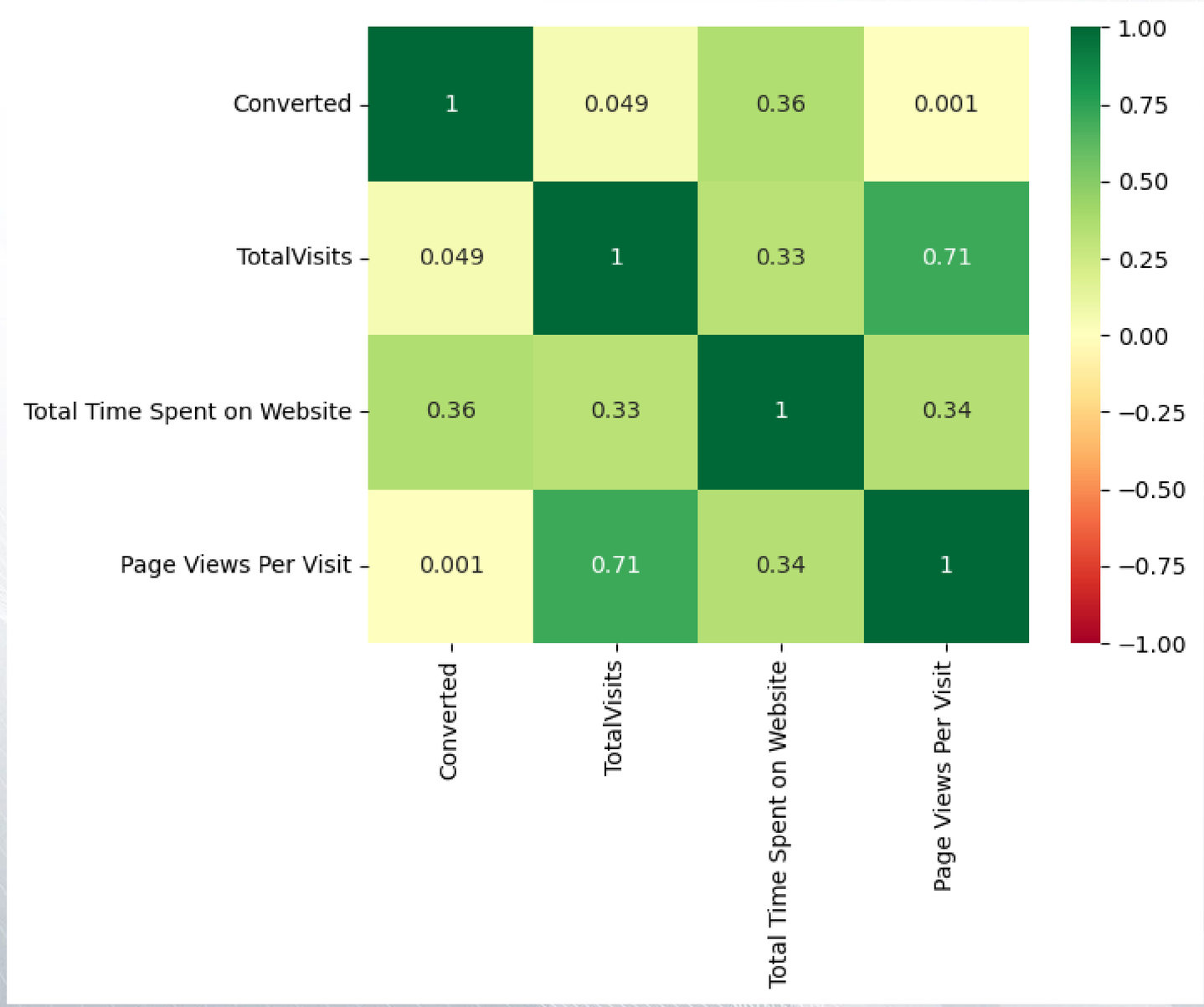
- Customers whose last notable activity was SMS Sent have a higher probability of conversion.



Bivariate Analysis

Correlation Heatmap

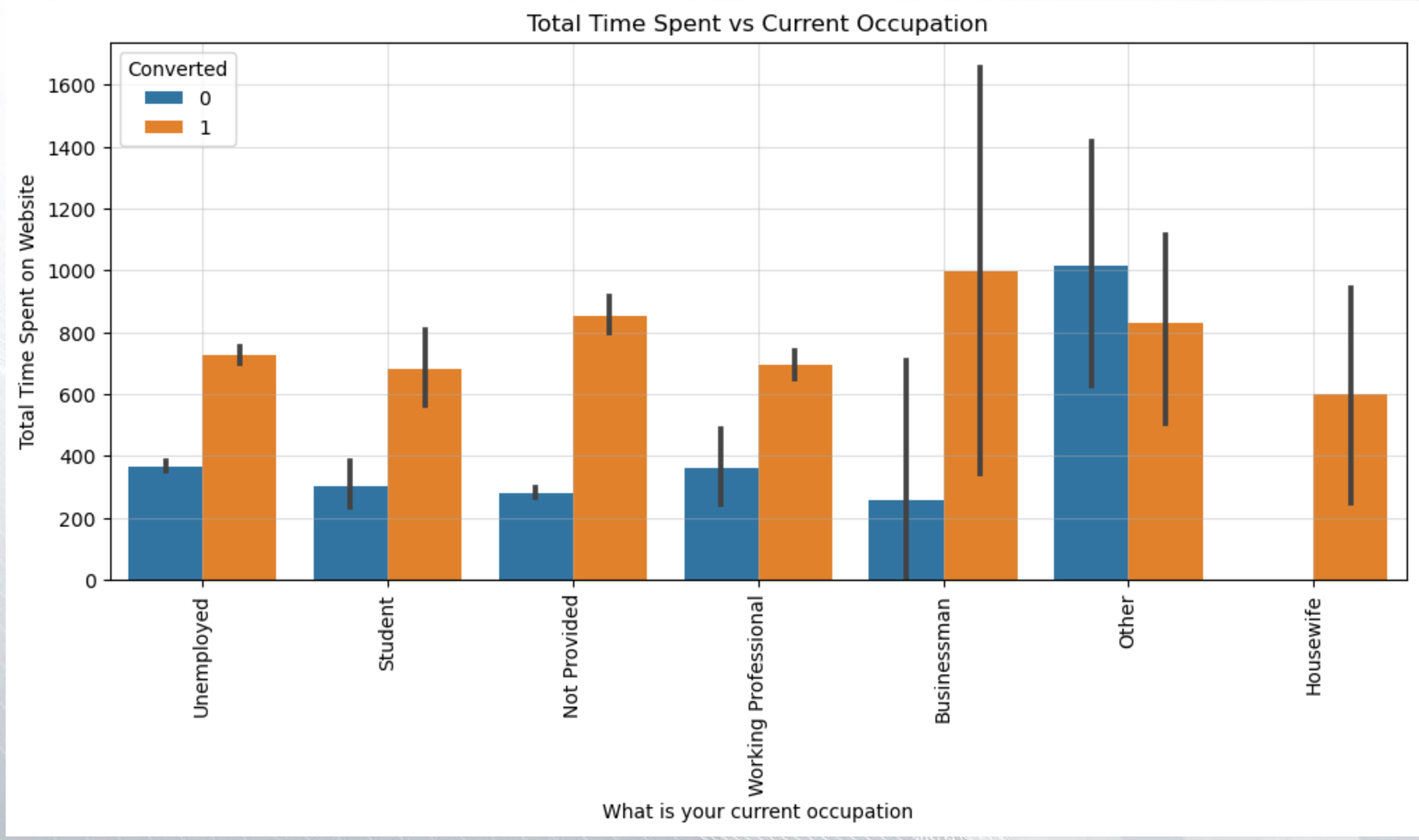
- We can see that TotalVisits and Page Views Per Visit have a good correlation among all the variables.



Bivariate Analysis

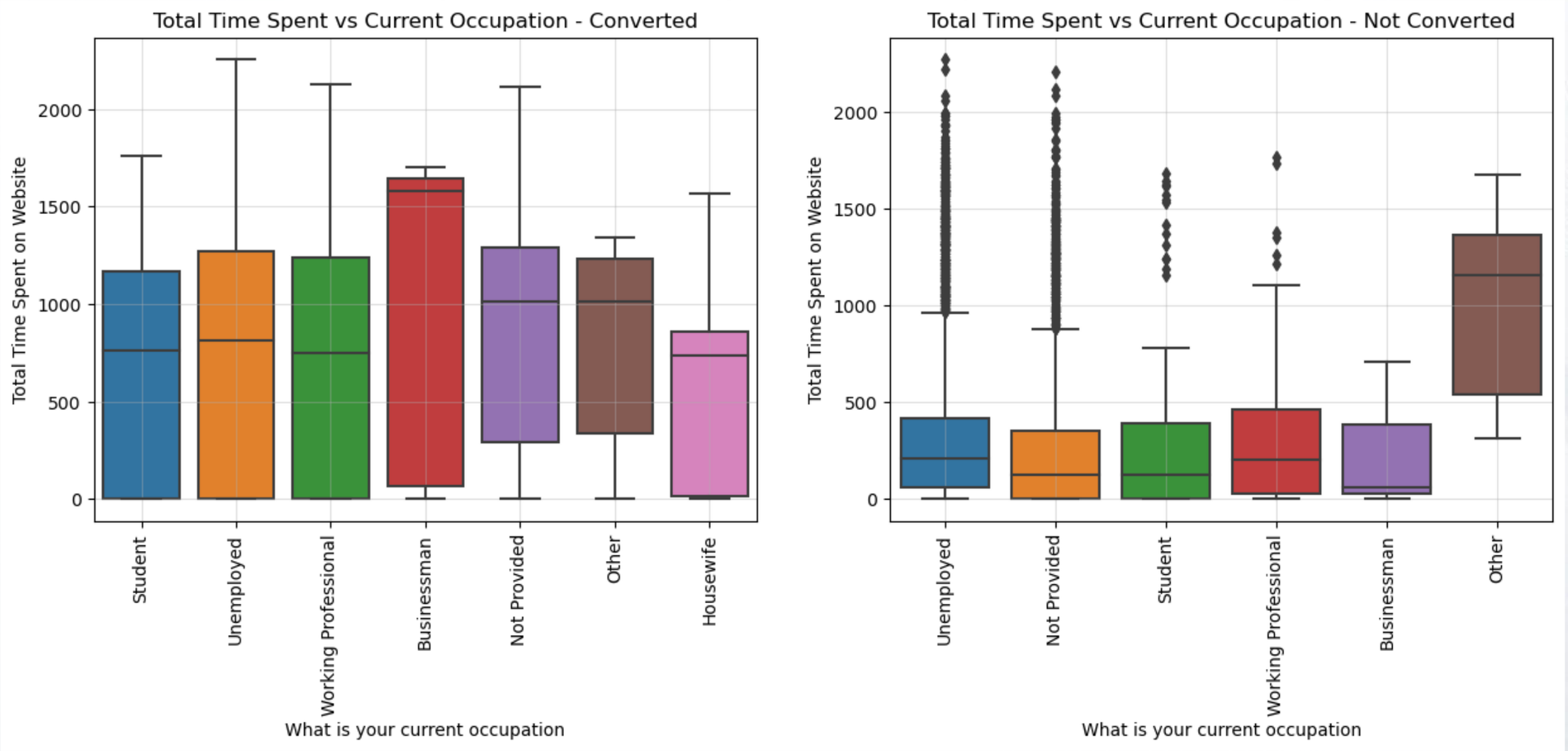
Total Time Spent vs Current Occupation

- We can see that the average time spent by converted leads is more than the average time spent by non-converted leads except 'Others'.



Bivariate Analysis

Total Time Spent vs Current Occupation



- We can see that the median time spent by converted leads is more than the 75th percentile time spent by non-converted leads except 'Others'.

Data Preparation

- Created dummy variables for categorical columns 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', and 'Last Notable Activity'.
- Split data into train-test set using scikit-learn library.
- The shape of the train set is (6351, 71), and the test set is (2723, 71).
- Numerical value normalised using MinMaxScaler.

Model Building

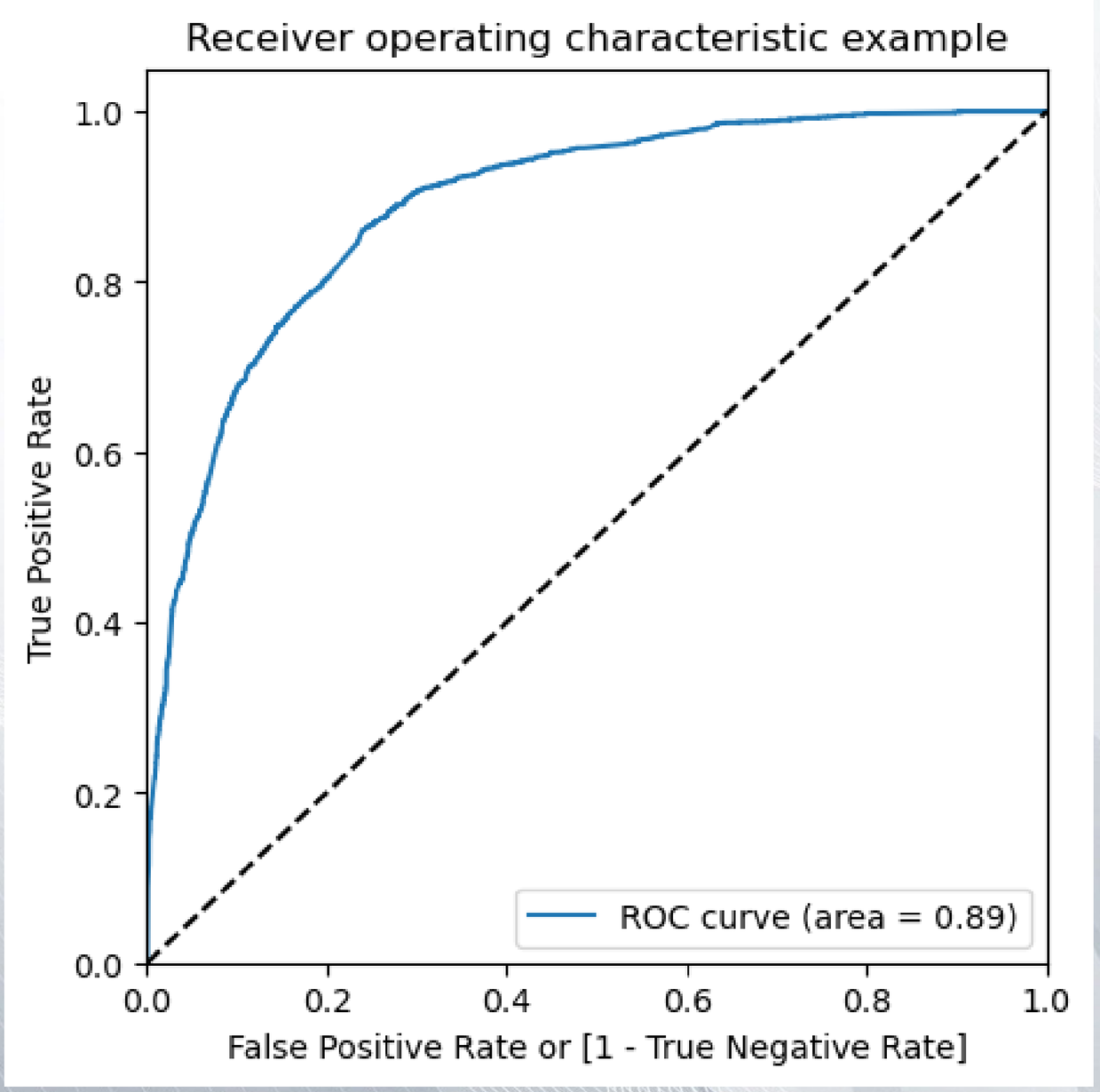
- Used RFE for selecting 15 features for model building.
- Eliminated feature with high p-value and vif value.
- Here are the features impacting the conversion rate.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.3352	0.107	-31.262	0.000	-3.544	-3.126
Do Not Email	-1.6841	0.171	-9.825	0.000	-2.020	-1.348
Total Time Spent on Website	4.6305	0.166	27.833	0.000	4.304	4.957
Lead Origin_Lead Add Form	3.6460	0.222	16.438	0.000	3.211	4.081
Lead Source_Olark Chat	1.4175	0.107	13.310	0.000	1.209	1.626
Lead Source_Welingak Website	1.9420	0.752	2.582	0.010	0.468	3.416
Last Activity_Converted to Lead	-1.3077	0.224	-5.836	0.000	-1.747	-0.869
Last Activity_Had a Phone Conversation	2.6959	0.746	3.614	0.000	1.234	4.158
Last Activity_Olark Chat Conversation	-1.3608	0.164	-8.307	0.000	-1.682	-1.040
Last Activity_SMS Sent	1.2317	0.076	16.311	0.000	1.084	1.380
What is your current occupation_Other	1.9403	0.712	2.727	0.006	0.546	3.335
What is your current occupation_Student	1.3623	0.230	5.928	0.000	0.912	1.813
What is your current occupation_Unemployed	1.2513	0.088	14.287	0.000	1.080	1.423
What is your current occupation_Working Professional	3.7472	0.199	18.811	0.000	3.357	4.138
Last Notable Activity_Unreachable	2.0238	0.490	4.127	0.000	1.063	2.985

Model Evaluation

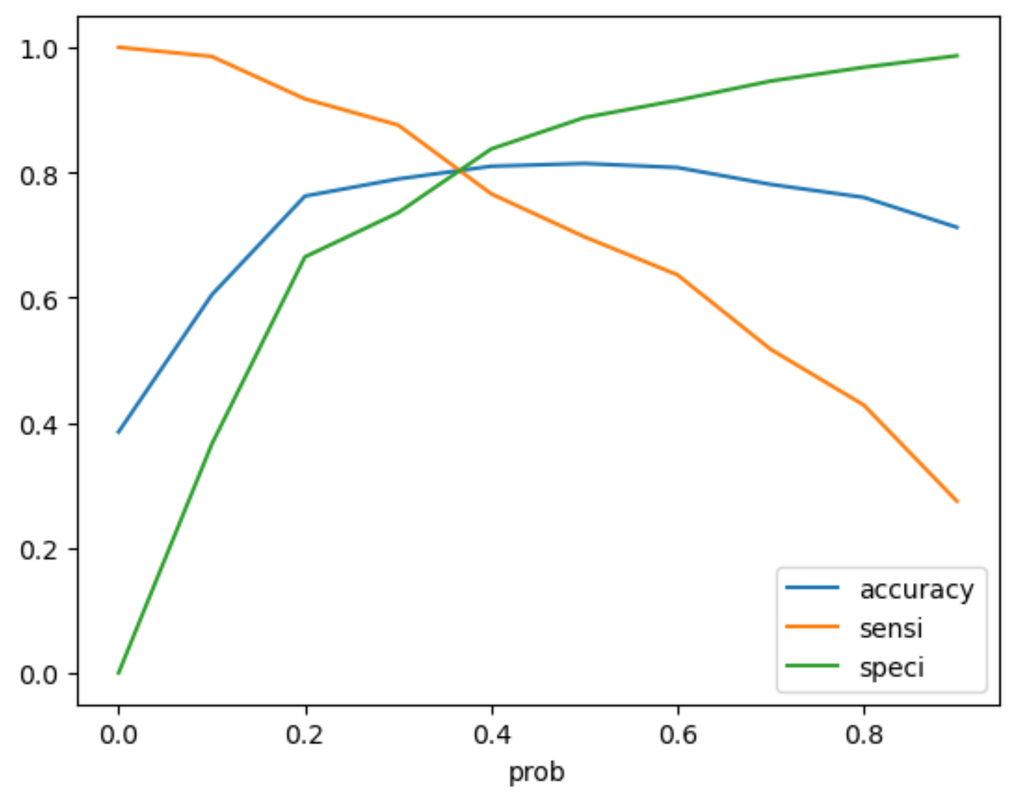
ROC Curve

- The area under the curve is 0.89 which shows that model is good.



Model Evaluation

Sensitivity and Specificity on Train Set



- The graph depicts 0.36 as the optimum point to take as a cutoff probability.

Confusion Matrix	Predicted Not Converted	Predicted Converted
Actual Not Converted	3160	745
Actual Converted	507	1939

- Accuracy - 80%
- Sensitivity - 79%
- Specificity - 81%
- False Positive Rate - 19%
- Positive Predictive Value - 72%
- Negative Predictive Value - 86%

Model Evaluation

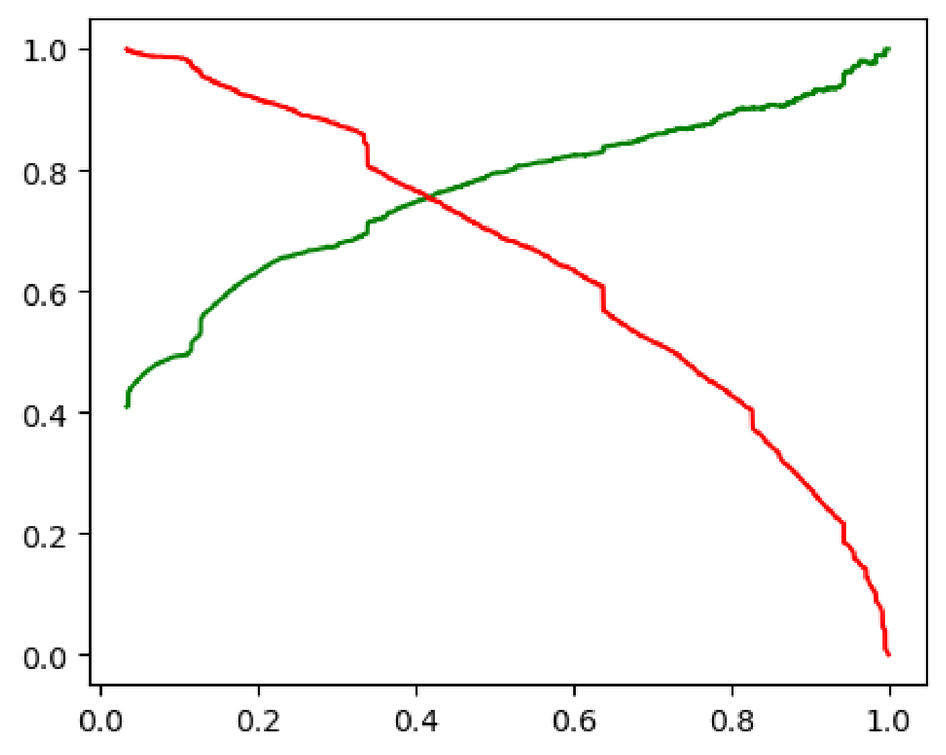
Sensitivity and Specificity on Test Set

Confusion Matrix	Predicted Not Converted	Predicted Converted
Actual Not Converted	1421	313
Actual Converted	215	774

- Accuracy - 81%
- Sensitivity - 78%
- Specificity - 82%
- False Positive Rate - 18%
- Positive Predictive Value - 71%
- Negative Predictive Value - 87%

Model Evaluation

Precision and Recall on Train Set



Confusion Matrix	Predicted Not Converted	Predicted Converted
Actual Not Converted	3313	592
Actual Converted	607	1839

- Accuracy - 81%
- Precision - 76%
- Recall - 75%

- The graph depicts 0.42 as the optimum point to take as a cutoff probability.

Model Evaluation

Precision and Recall on Train Set

Confusion Matrix	Predicted Not Converted	Predicted Converted
Actual Not Converted	1486	248
Actual Converted	248	741

- Accuracy - 82%
- Precision - 75%
- Recall - 75%

Conclusion

Top 3 variables contribute most in lead conversion:

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional

Top 3 categorical/dummy variables should be focused most for lead conversion:

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Last Activity_Had a Phone Conversation

Conclusion

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Answer: Calls should be made to those people if:

- They are working professionals spending 700 or more time on the website are potential leads. These people should be priorities.
- Lead source is Welingak_Website, Google, and Reference, as they have a higher conversion rate.
- Lead source is Olark_chat
- Lead origin is Lead Add Form and there last activity is Had a Phone Conversation

Conclusion

Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Answer: During this time sales team should make calls to only those customers whose lead score is above 70 to avoid useless phone calls and also take feedback from existing customers using Typeforms and pass it to the business/analytics team for product and service improvement.

Recommendations

- People spending more time on the website have a good chance of buying the course.
- Demand for management courses is more than for other courses.
- Leads from Working professionals have a good chance of converting.
- When a lead origin is Lead Add Form and the last activity is Had a phone conversation mostly tends to be converted.

X Education should closely monitor if a lead is coming from the below sources because they have a good conversion rate or they contribute most to lead conversion:

- Welingak_website
- Olark Chat
- Referral
- Google

THANK YOU