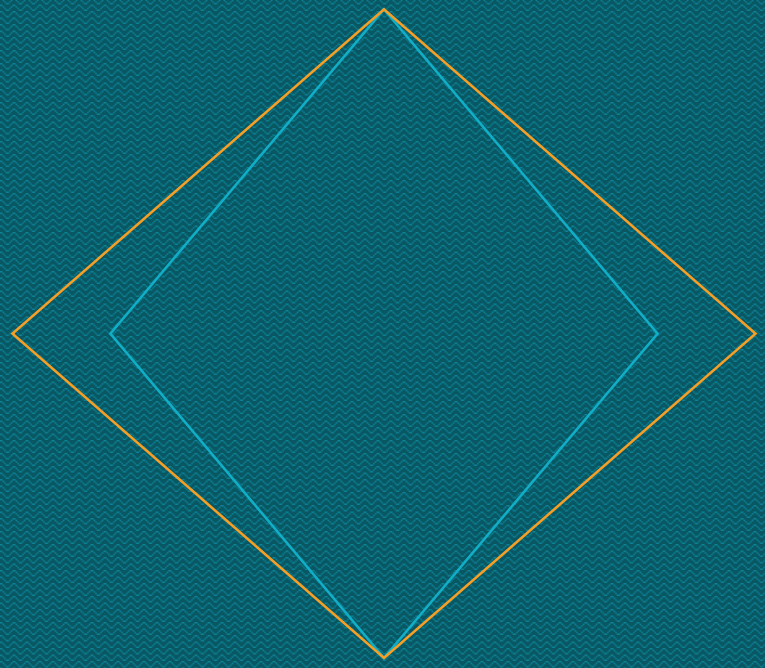


Predictive analytics for X-Education

Vivek Sharma

EPGP – Data Science, IIIT, Bangalore

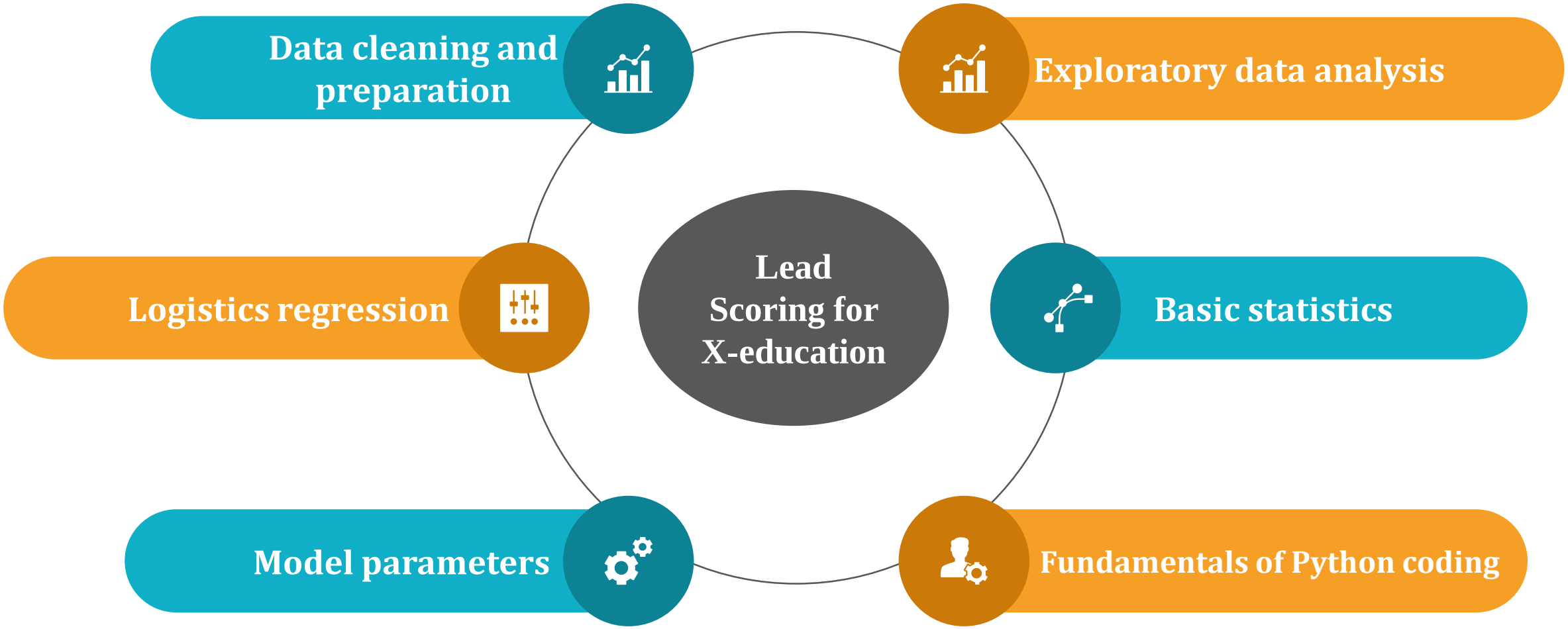
Problem statement and offered solution

Client statement: X Education appointed me to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company required a build a model wherein we assigned 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 gave an estimated target lead conversion rate to be around 80%.

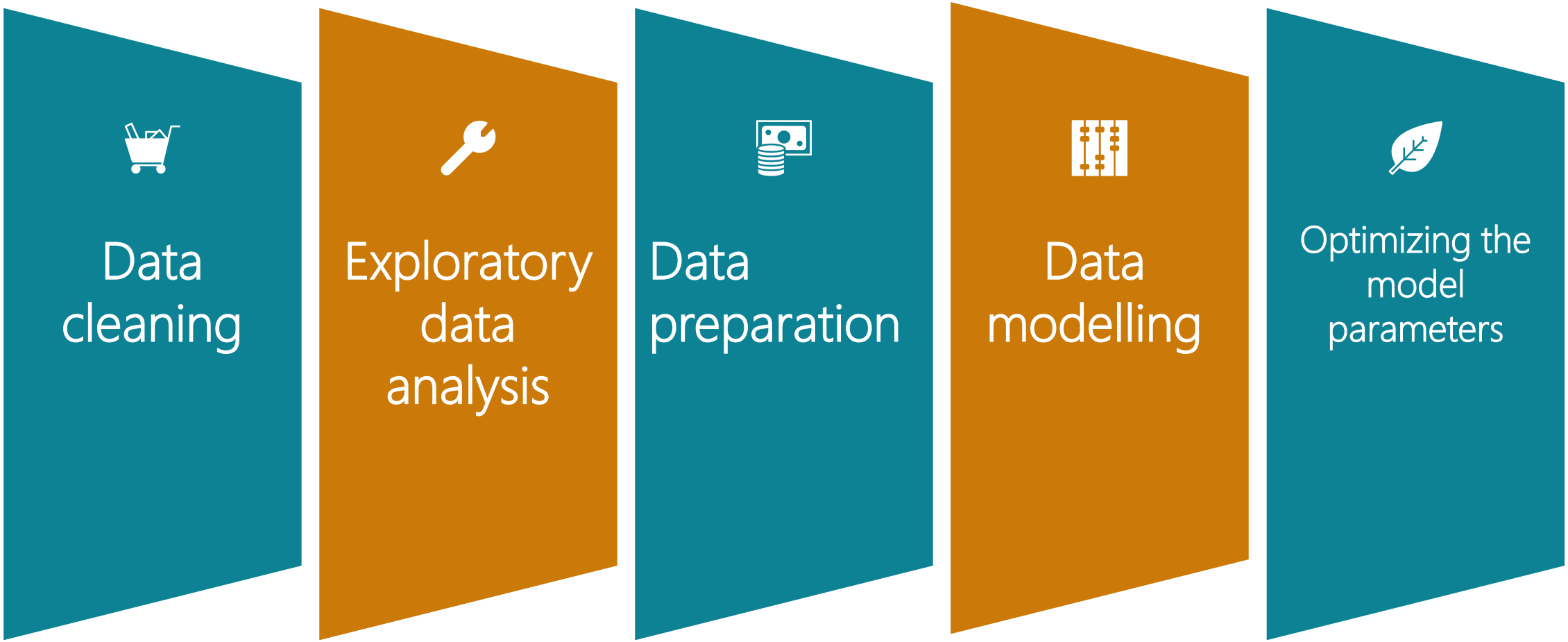
Offered solution: I did the cleaning of the data bringing focus to the variables that will really help in making predictions and in the same pursuit, variables with missing values > 30% were dropped off, followed by logical imputation of missing values for few variables and merging few that contained similar nature of data. Further to this, the exploratory data analysis was performed, which gave us a clearer idea about impact of many variables on the target variable (Lead conversion).

Next to this, predictive modelling was done for achieving the goal set-up by client through Logistic regression models targeted towards high accuracy, precision, sensitivity and specificity which results in accurate predictions about lead conversion and can help X-education to achieve the said target.

TOOLS EMPLOYED



STRATEGY TO DRAW CONCLUSIONS

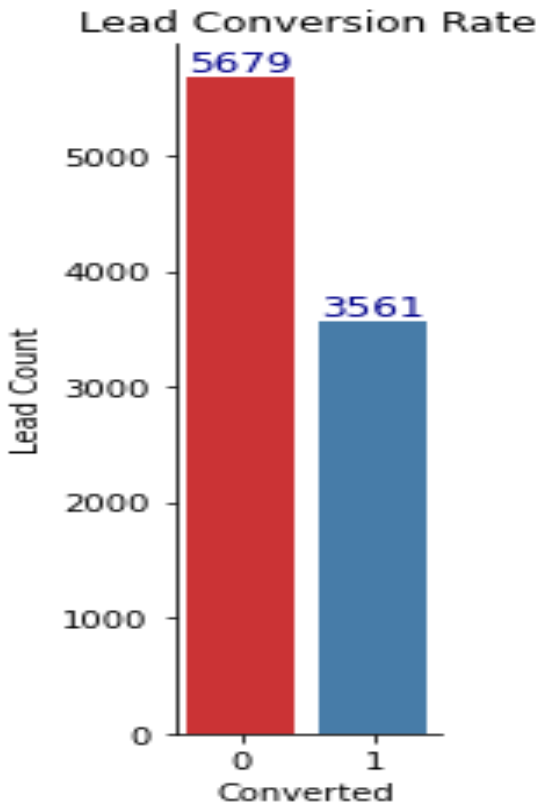


Exploratory data analysis

Lead Conversion Rate

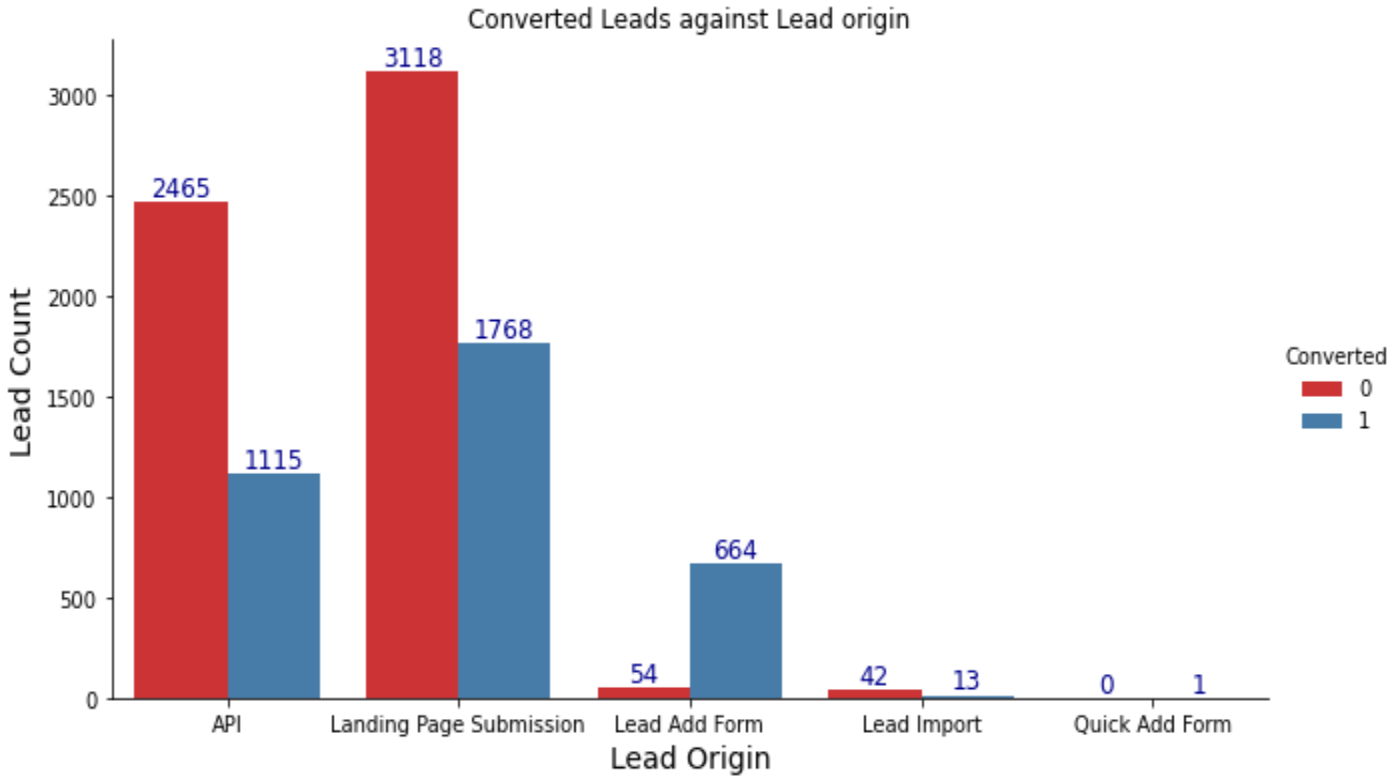
- The available data of target variable "Converted" shows the rate of conversion as = 39% approximately.

  Plot above is the representative of Converted and Not-converted (x-axis) verus Lead counts (y-axis).



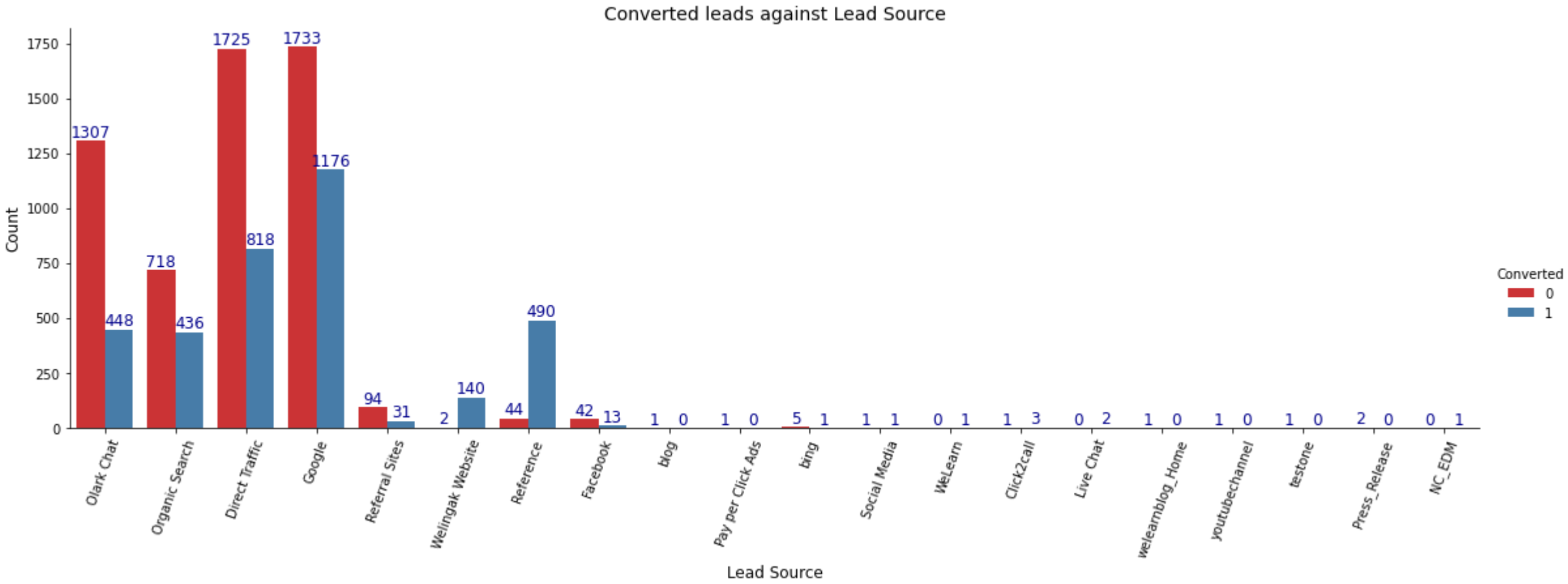
Converted leads vs Lead origin

Observation: The bar-plot above indicates that the maximum conversion came from the 'Landing Page Submission' followed by 'API', and then trailed by 'Lead Add Form'.



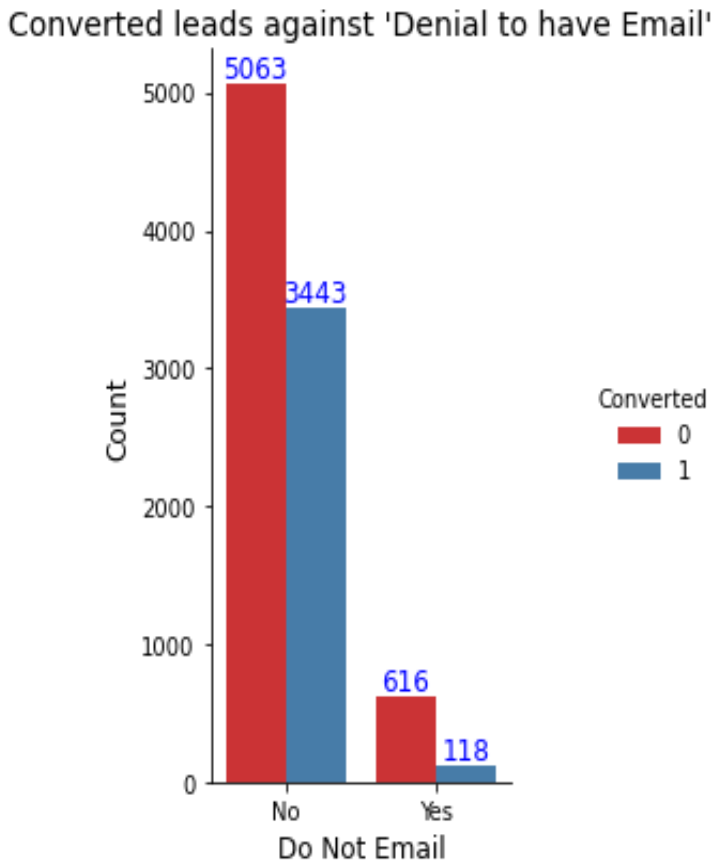
Converted leads vs Lead source

- Observation: Highest conversion came from 'Google', followed by 'Direct Traffic', further trailed by 'References', 'Olark Chat', and 'Organic Search'.



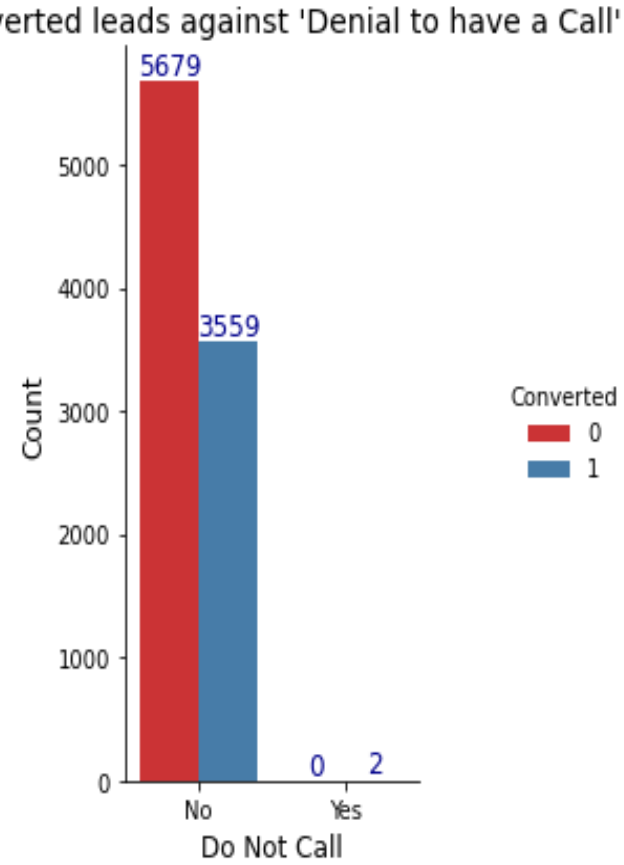
Converted leads vs Customers not wanting emails

- Observations: The potential leads who denied to have email from X-education hardly converted to actual leads (still around = 16%). Customers who accepted to have the email were converted in higher numbers.



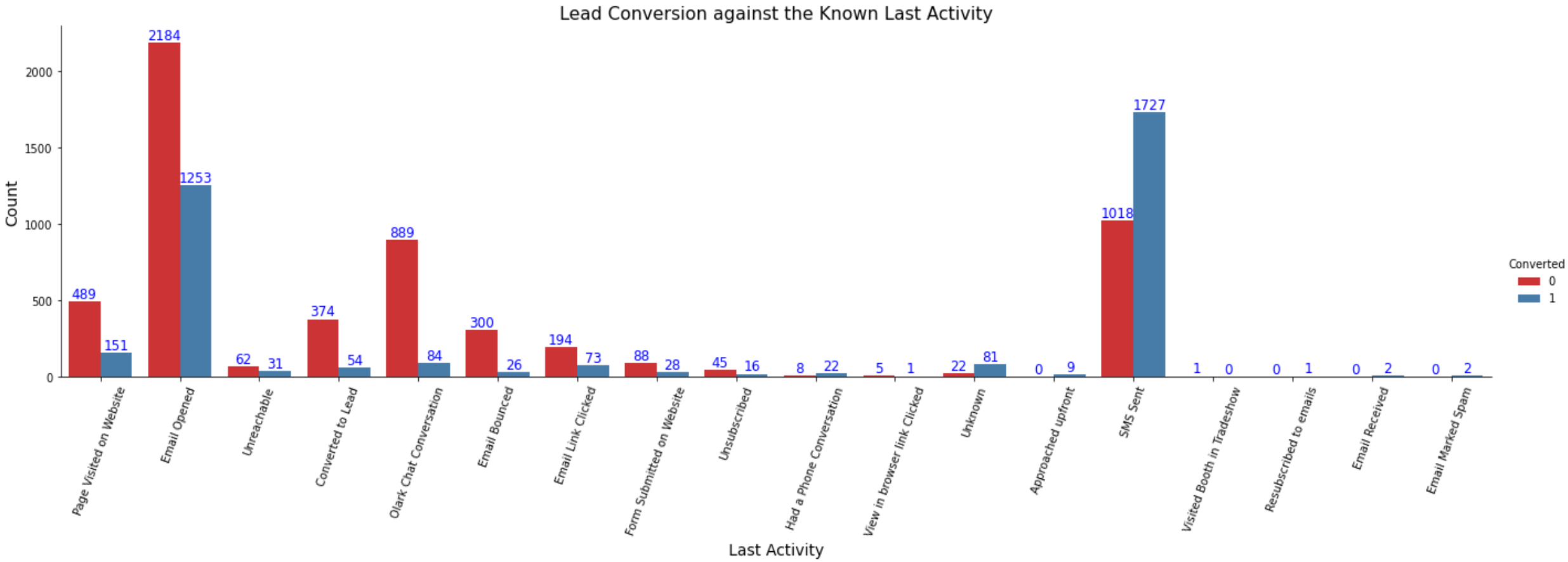
Converted leads vs Customers not wanting to get calls

- Observation: High number of conversions happened when Customers choose to receive a call from X-education. However, we can see here 2 leads getting converted even when they chose not to get a call.



Lead Conversion vs Known last activity

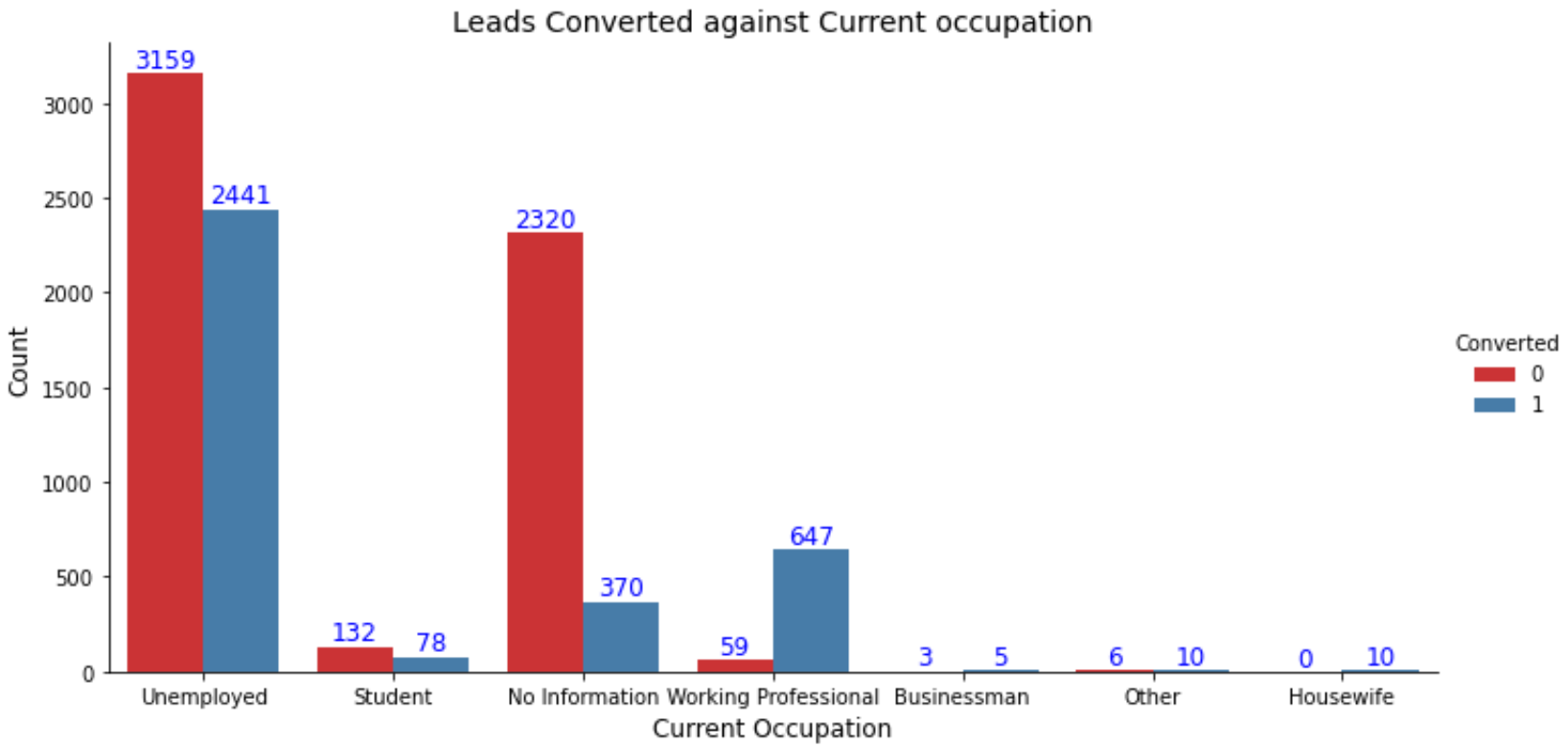
- Observation: It is observed here that when the customers responded last with SMS, that lead to highest conversion, followed by when they opened Email, further trailed by when they responded to Olark Chat and Visited Pages on website.



Leads conversion vs Current occupation

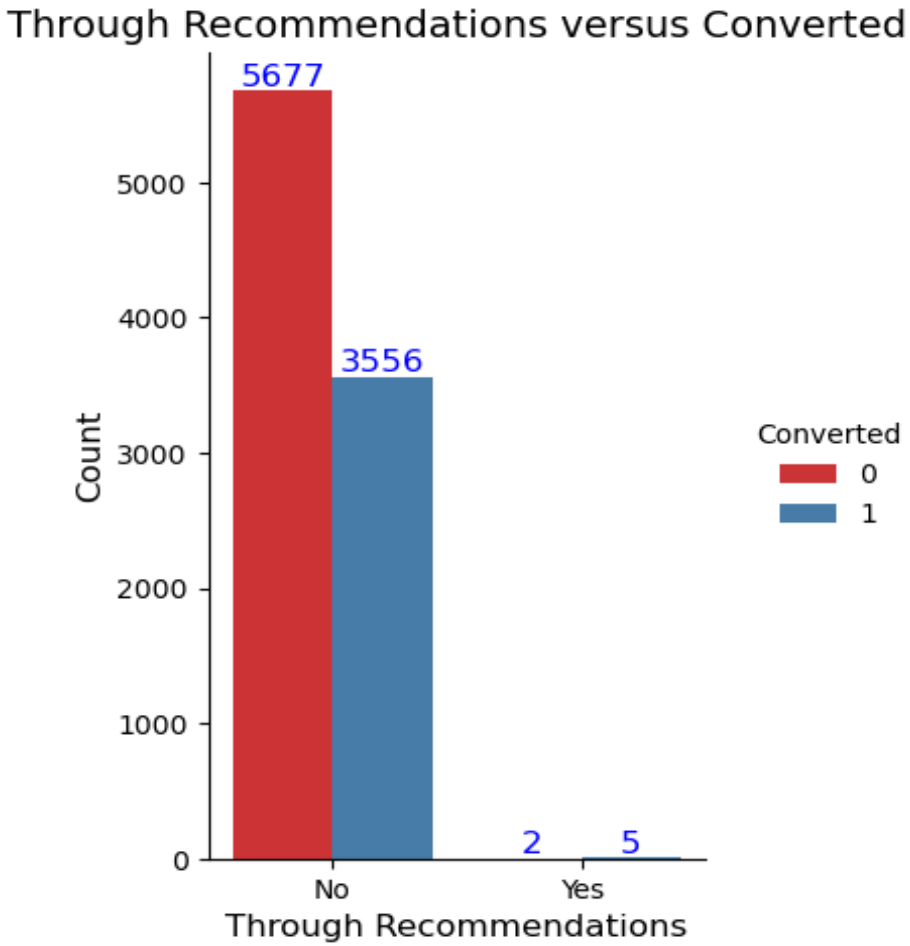
- Observation: Higher number of customers who were unemployed coverted into leads, followed by Working professionals and trailed by people who did not shared the occupation information, followed by students.

- Another striking observation is that out of 8 businessman 5 got converted, and 10 housewives approached who got converted at 100% rate.



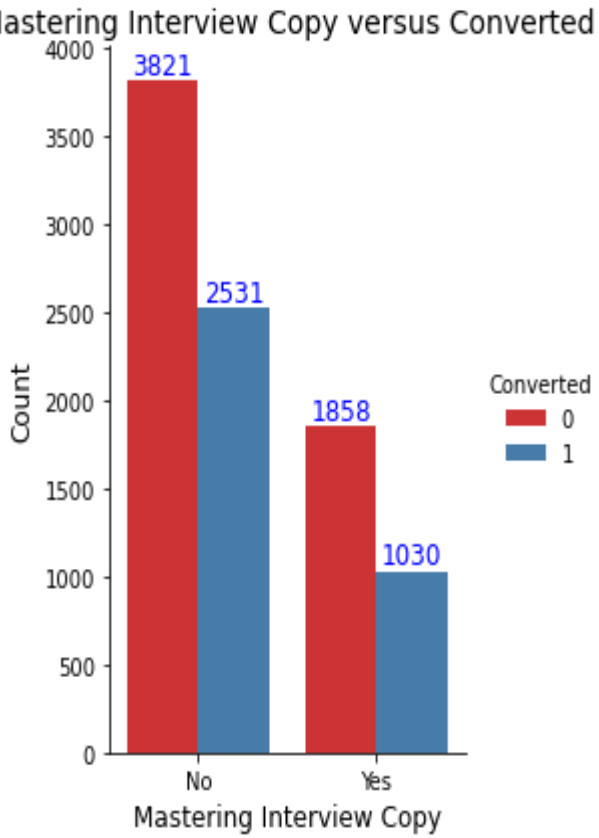
Leads conversion vs Recommendations

- Observations: The conversion through recommendations is having a significantly high number



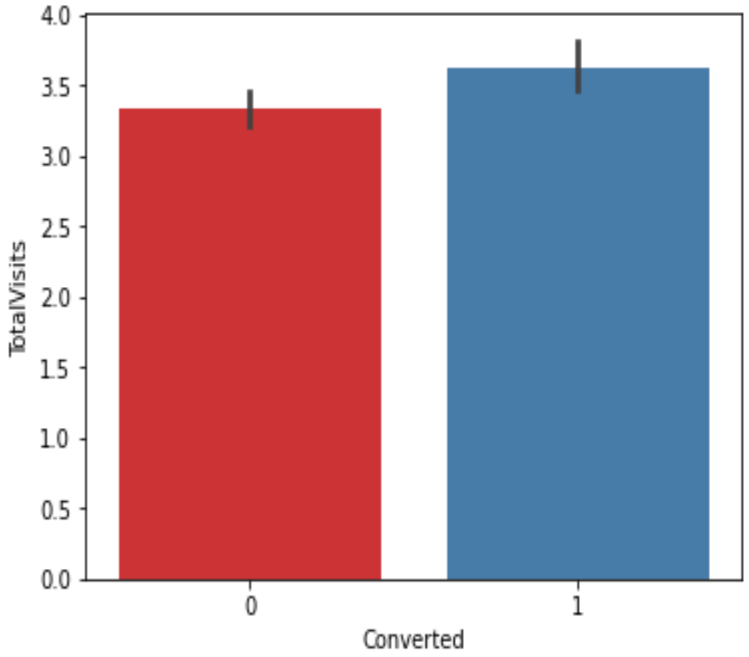
Insights based on Employment years of the client

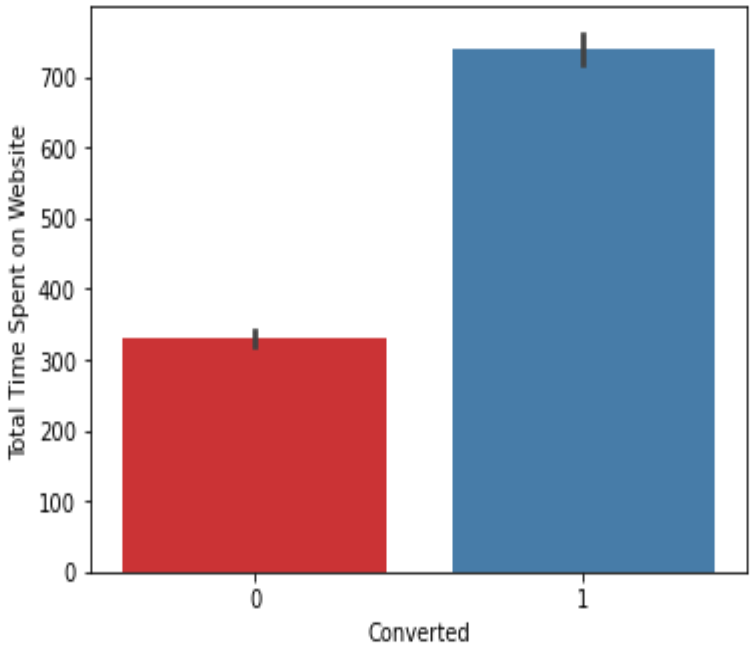
- Observations: Customers who asked for ‘Mastering Interview copy’ have converted well (around 35%), however the customers those did not asked for this material also showed conversion (around 39%).
- It means that it is not a very strong indicator, but company should continue to provide this material to its customers, so it engages them for small tome with the company atleast for a short period of time.

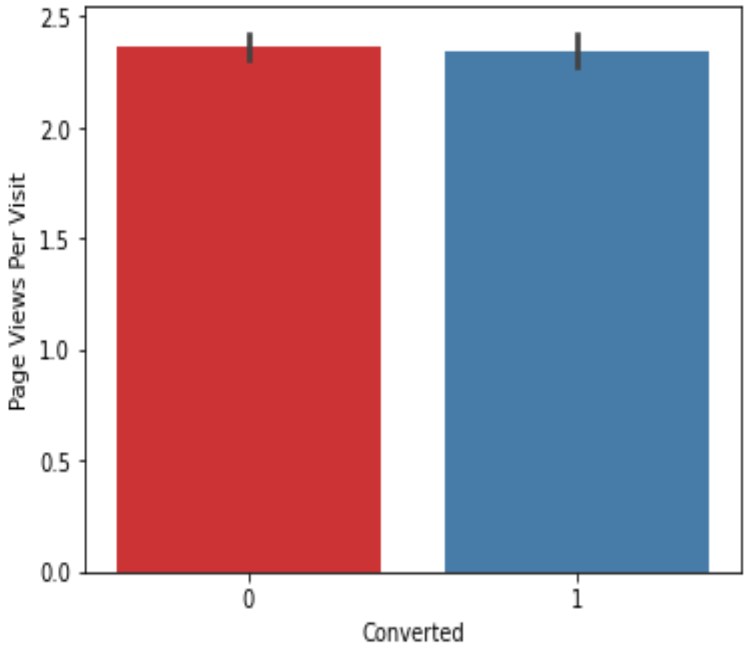


Conversion vs Total Visits, Total Time spent and Page views

- The car graph below indicating that Total time spent on the website of X-education and the page views alongside total visits are very string indicators of lead conversion. Hence company should continue focus on the quality of their website and keep the potential clients engaged with an intuitive and content rich website.







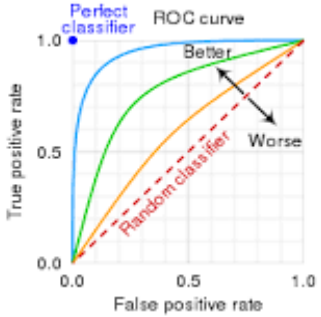
Predictive model building and evaluation

Best suited Logistics regression model

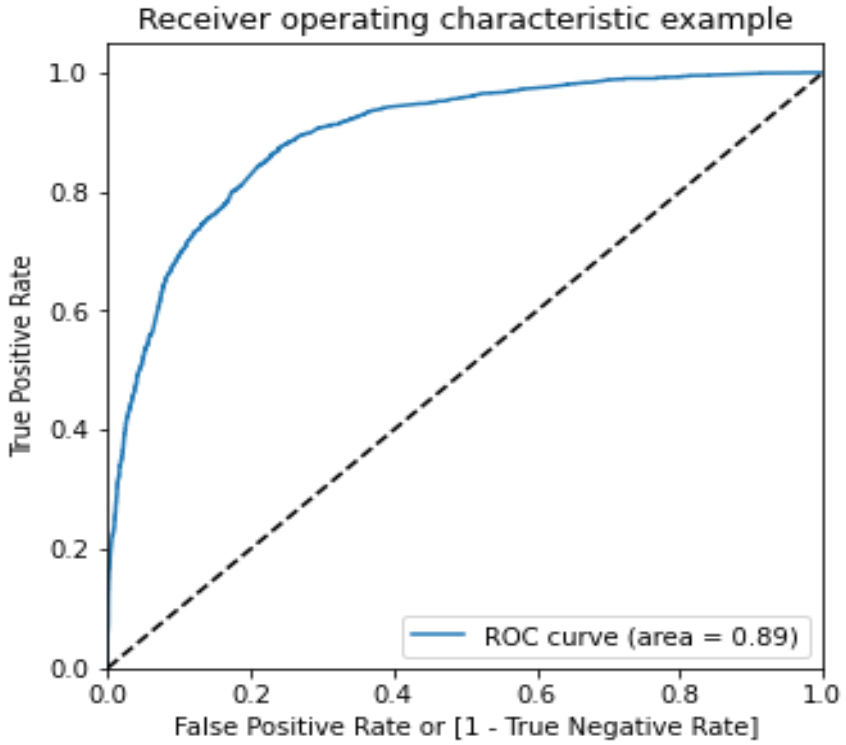
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6446
Model Family:	Binomial	Df Model:	21
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2573.4
Date:	Tue, 21 Mar 2023	Deviance:	5146.7
Time:	12:16:40	Pearson chi2:	6.88e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.4140
Covariance Type:	nonrobust		

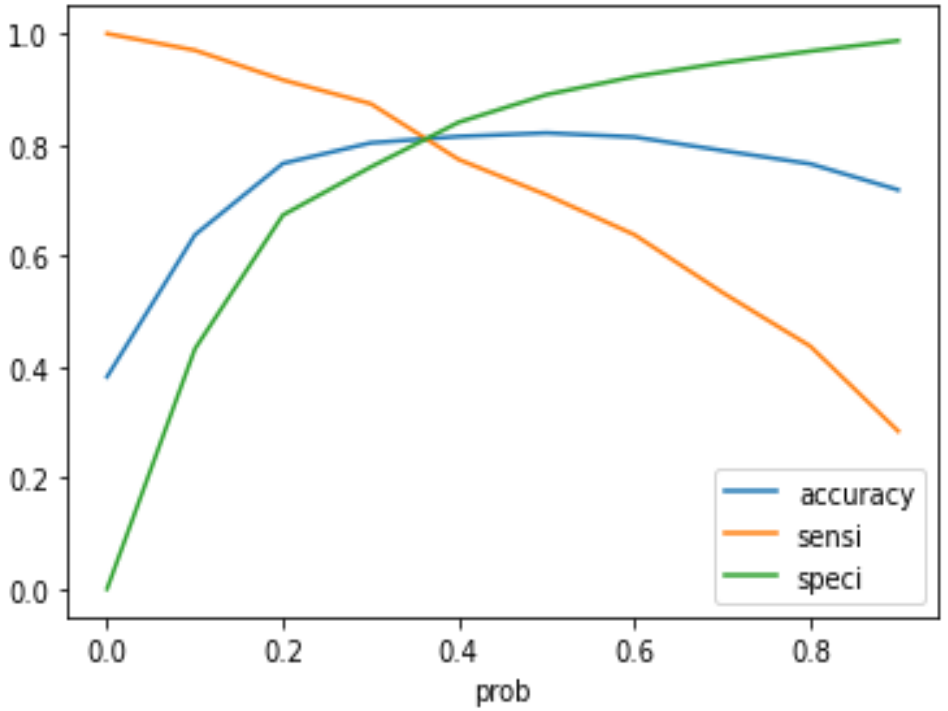
Attributes of best suited logistics regression model



How an ideal ROC will be?



ROC of the best suited model



Deciding a cut-off for the data, with focus on accuracy, sensitivity and specificity

Best suited Logistics regression model

Total 5 Logistics regression models were made, out of which 5th was selected based on following parameters for training data set:

1. We can see that the final prediction of conversions have a target of 80% (79.8%) conversion as per the X Educations CEO's requirement, which is achieved by 5th Model. Hence this is a good model to be used by X-education.
2. Sensitivity = 70%
3. Specificity = 89%
4. Accuracy = 81%

Best suited Logistics regression model

The selected 5th model trained on the ‘train data set’ is used for prediction on the ‘test data’ and the following parameters for test data set are indicative of its relevance to the current business problem.

1. Accuracy of the test data set as per 5th Logistics regression model is 81.67%
2. Sensitivity of the test data set as per the 5th Logistics regression model is 78.89 % ~ 79%
3. Specificity of the test data set as per the 5th Logistics regression model is 83.47%

Hence the 5th Logistics regression model is suitable for making predictions on the current data which can help X-education company to increase their lead conversion and churn out good business.


Thank You