Lead Scoring

X Education

From:

Meenali Sharma

Objective

Build Logistic regression model & assign Lead Scores to prospective candidates of X Education

Problem description

X Education is an online Education company

- It has Lead database, some of which get converted & some not
- The typical lead conversion rate is 30% which needs to be maximized to at least 80%
- Target is to identify Hot leads which have a high conversion rate
- The hot leads to be identified by cutoff Lead Scores
- Lead scores to be assigned to each candidates based on probabilities calculated by Logistic regression model

Contents

- Data- Raw, Preparation, Cleaned
- Exploratory data analysis- Results
- Dummy variable creation
- Logistic regression modelling
- Model performance evaluation-ROC curve, accuracy
- Model fit on test data
- Conversion results- percentage at different Lead score cut offs
- Conclusion
- Recommendations

Data- Raw

Prospect ID	Specialization	Tags
Lead Number	How did you hear about X Education	Lead Quality
Lead Origin	What is your current occupation	Update me on Supply Chain Content
Lead Source	What matters most to you in choosing this course	Get updates on DM Content
Do Not Email	Search	Lead Profile
Do Not Call	Magazine	City
Converted	Newspaper Article	Asymmetrique Activity Index
TotalVisits	X Education Forums	Asymmetrique Profile Index
Total Time Spent on Website	Newspaper	Asymmetrique Activity Score
Page Views Per Visit	Digital Advertisement	Asymmetrique Profile Score
Last Activity	Through Recommendations	I agree to pay the amount through cheque
Country	Receive More Updates About Our Courses	Last Notable Activity
a free copy of Mastering The Interview		

Raw dataset contains 9240 rows and 37 columns.

Data Preparation

Missing value treatment

- Column containing >70% missing data were dropped.
- Asymmetrique Index columns were checked for any possible relation to impute missing values
- In absence of any visible correlation with Activity & Profile, these columns were dropped too
- ‘City’ column had ~40% missing values & was dropped
- Other columns with possible imputations were handled appropriately

Unique value columns

- Columns with only one type of unique values were dropped in absence of variability

Imputation

- High missing value containing columns were imputed with suitable values

Cleaned dataset

Lead Origin	Lead Source
Do Not Email	Converted
Total Visits	Total Time Spent on Website
Page Views Per Visit	Last Activity
Country	Specialization
Tags	Lead Quality
What is your current occupation	
A free copy of Mastering The Interview	
Last Notable Activity	

- After data cleaning, 15 column are left.

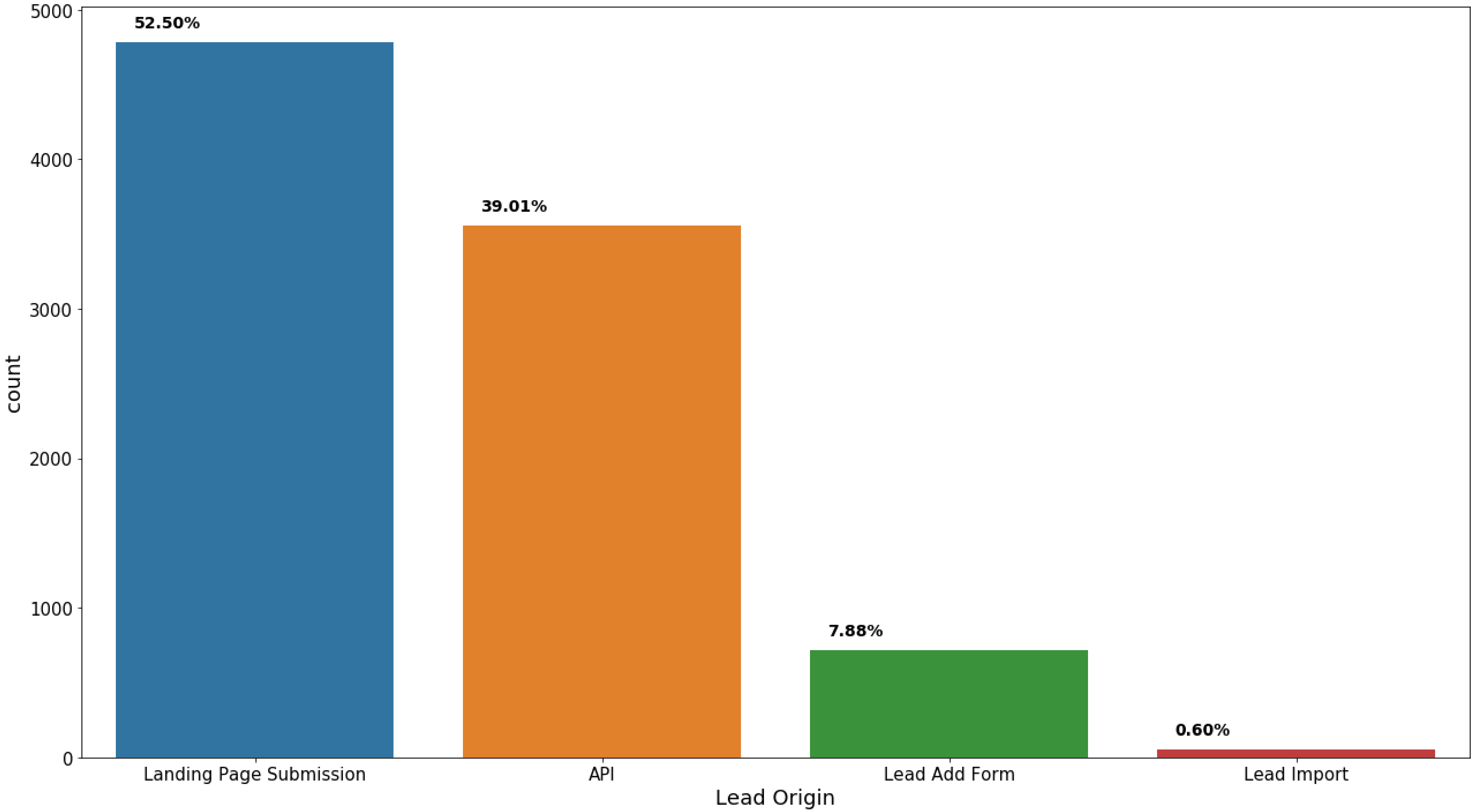
Outlier treatment

- Numeric columns were treated for Outliers
- Data within +/- 3*Standard deviation was retained

Exploratory Data Analysis

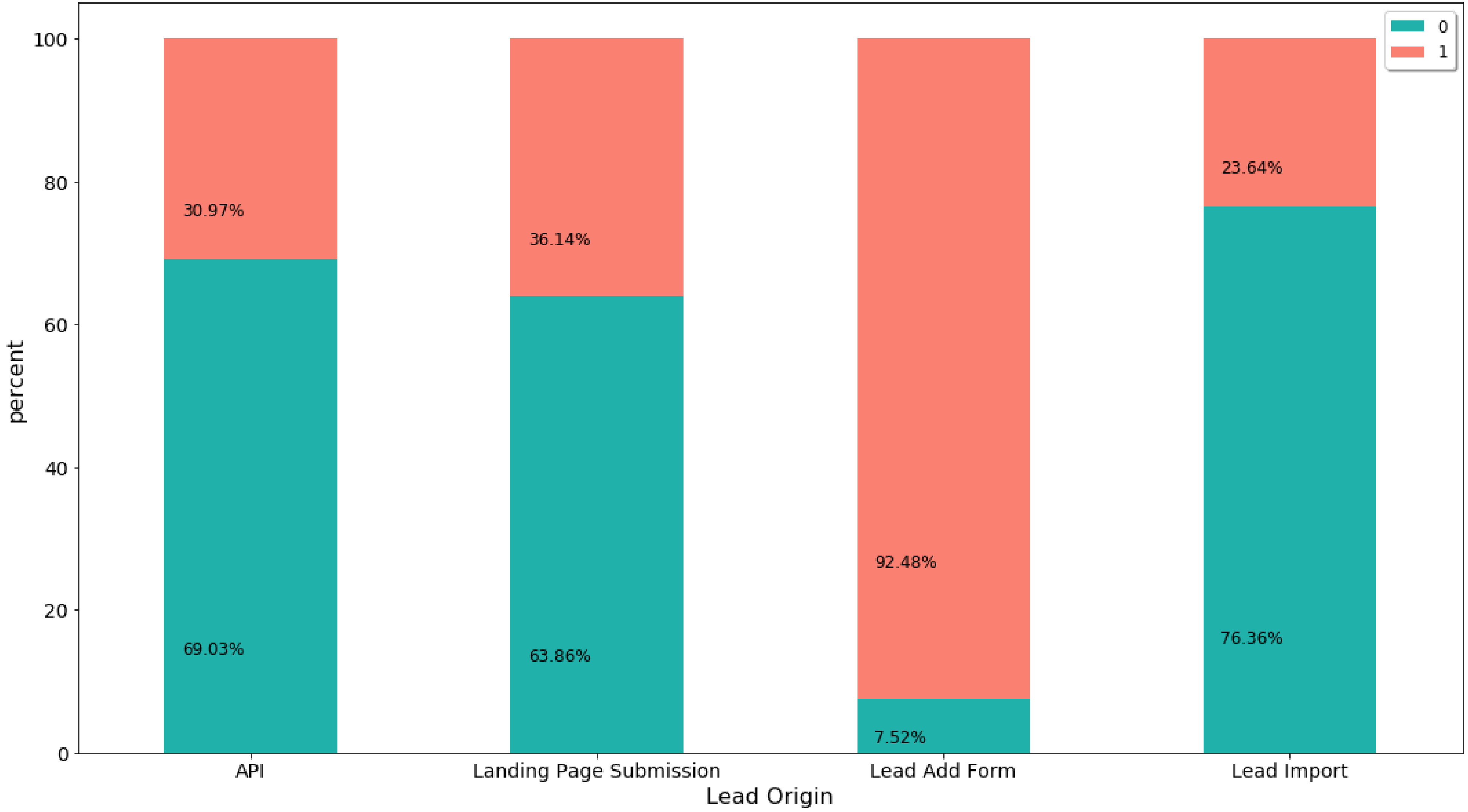
- Univariate and Segmented Univariate analysis was performed for each column variable
- The analysis is done with respect to Conversion
- Univariate graphs display a particular variables' individual class contribution in entire dataset
- Segmented Univariate analysis displays Conversion rate across each class
- EDA is helpful to understand class requiring focus under each variable
- Prominent results of EDA are presented-

Lead Origin



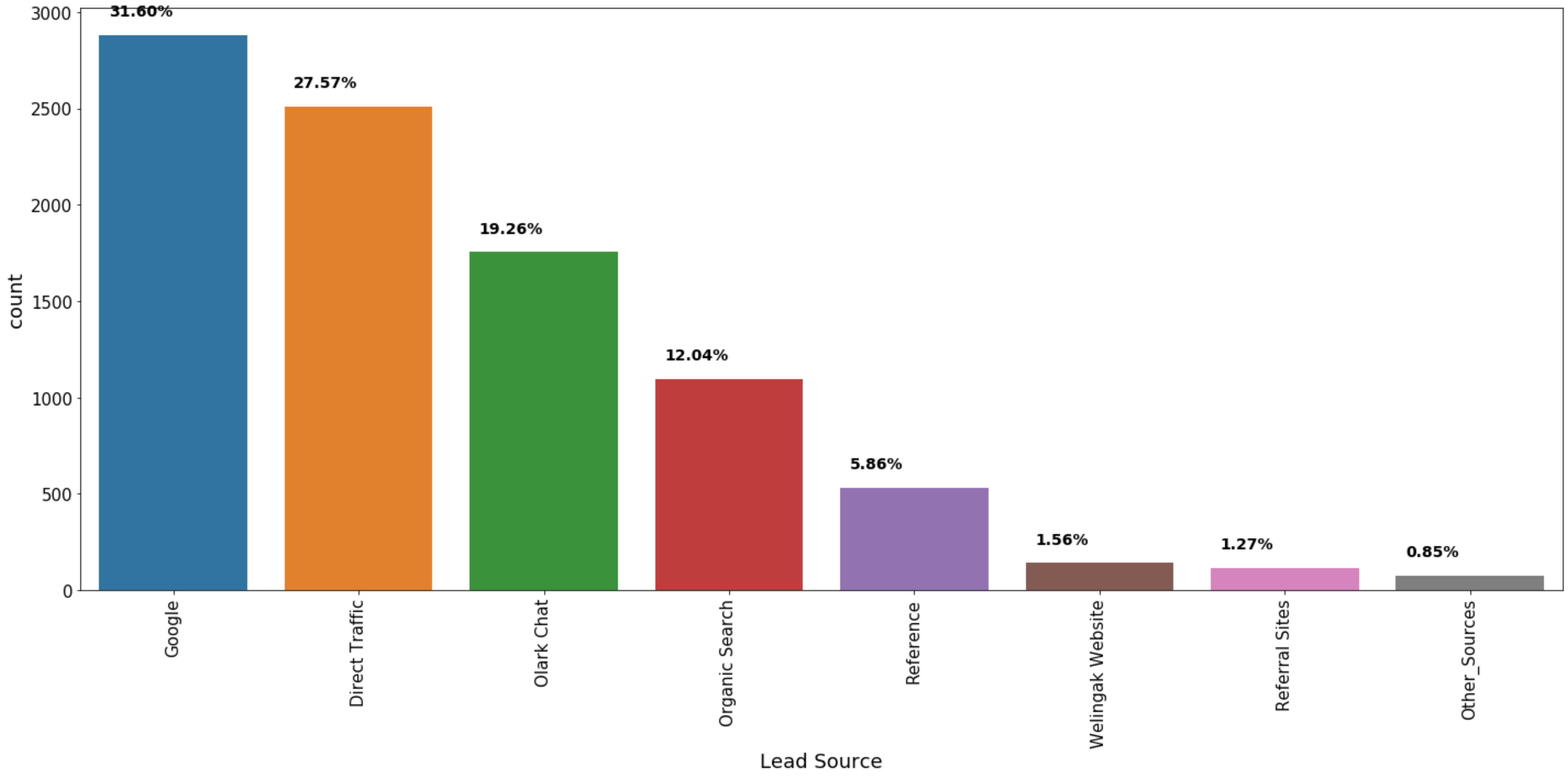
Order of Lead origin (Max to Min)- Landing Page Submission -> API -> Lead Add Form -> Lead Import

Lead Origin



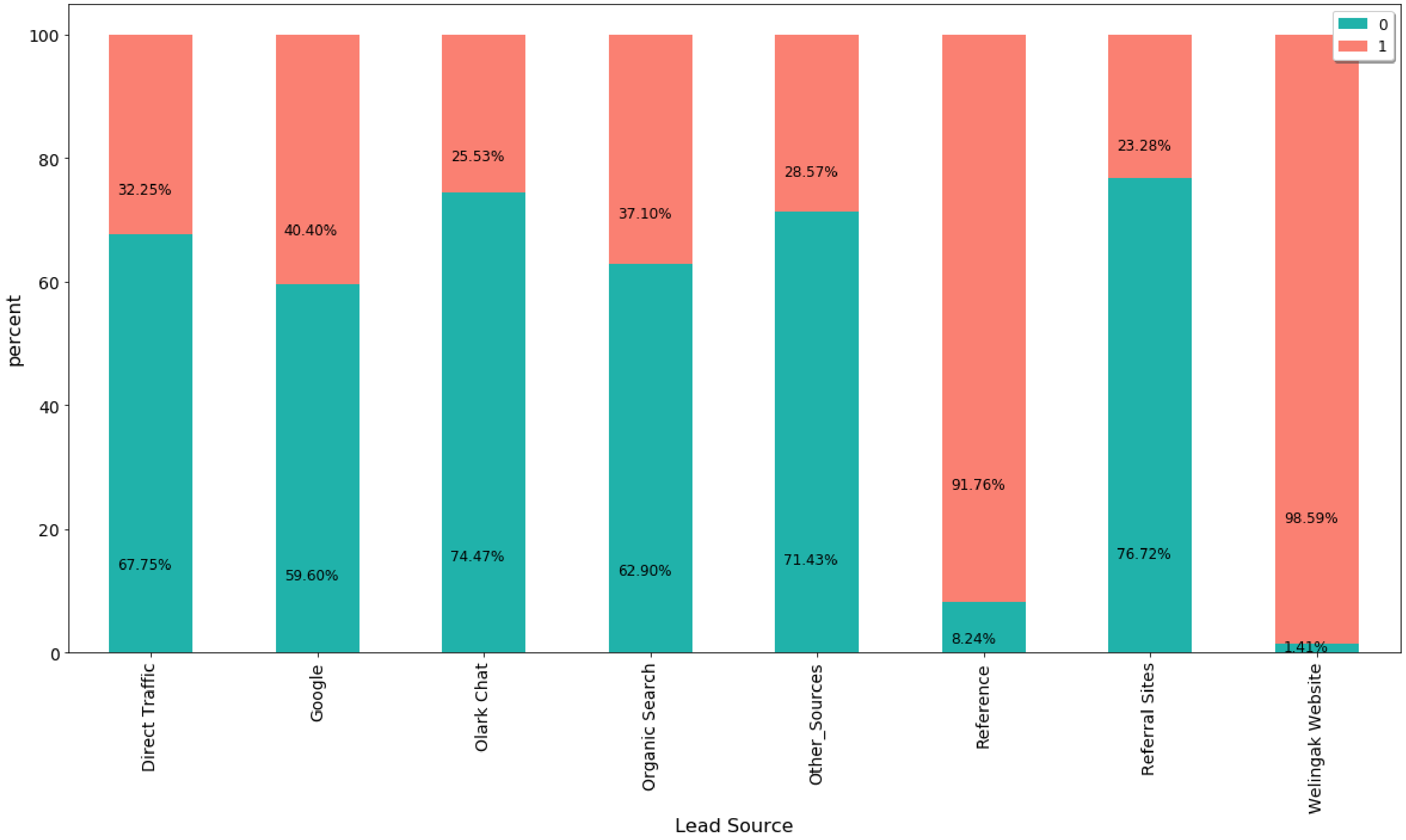
- Order of Conversion rate (Max to Min)-Lead Add Form -> Landing Page Submission -> API -> Lead Import.
- Company should focus on High conversion providing "Lead ADD Form"

Lead Source



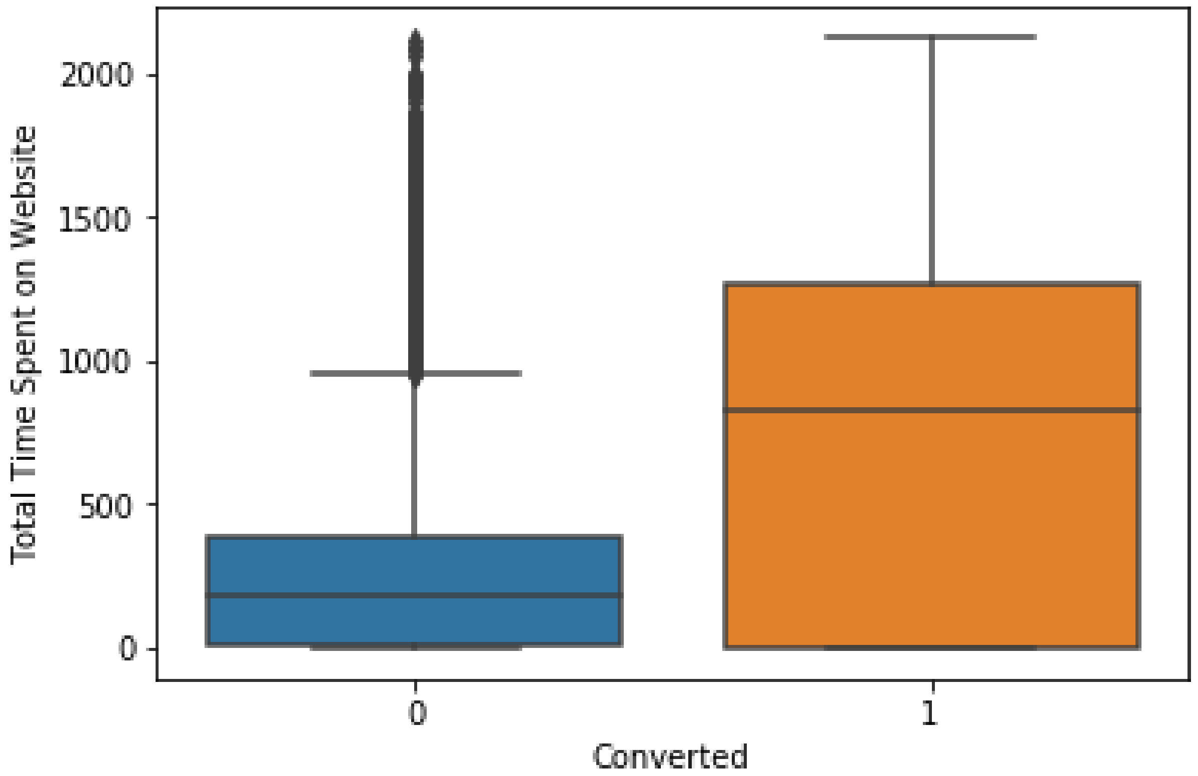
- 'Lead Source' order (Max to Min)- Google -> Direct Traffic -> Olark Chat -> Organic Search -> Reference -> Welingkar Website -> Referral Sites -> Other_Sources
- Maximum 32% Leads are coming from Google

Lead Source



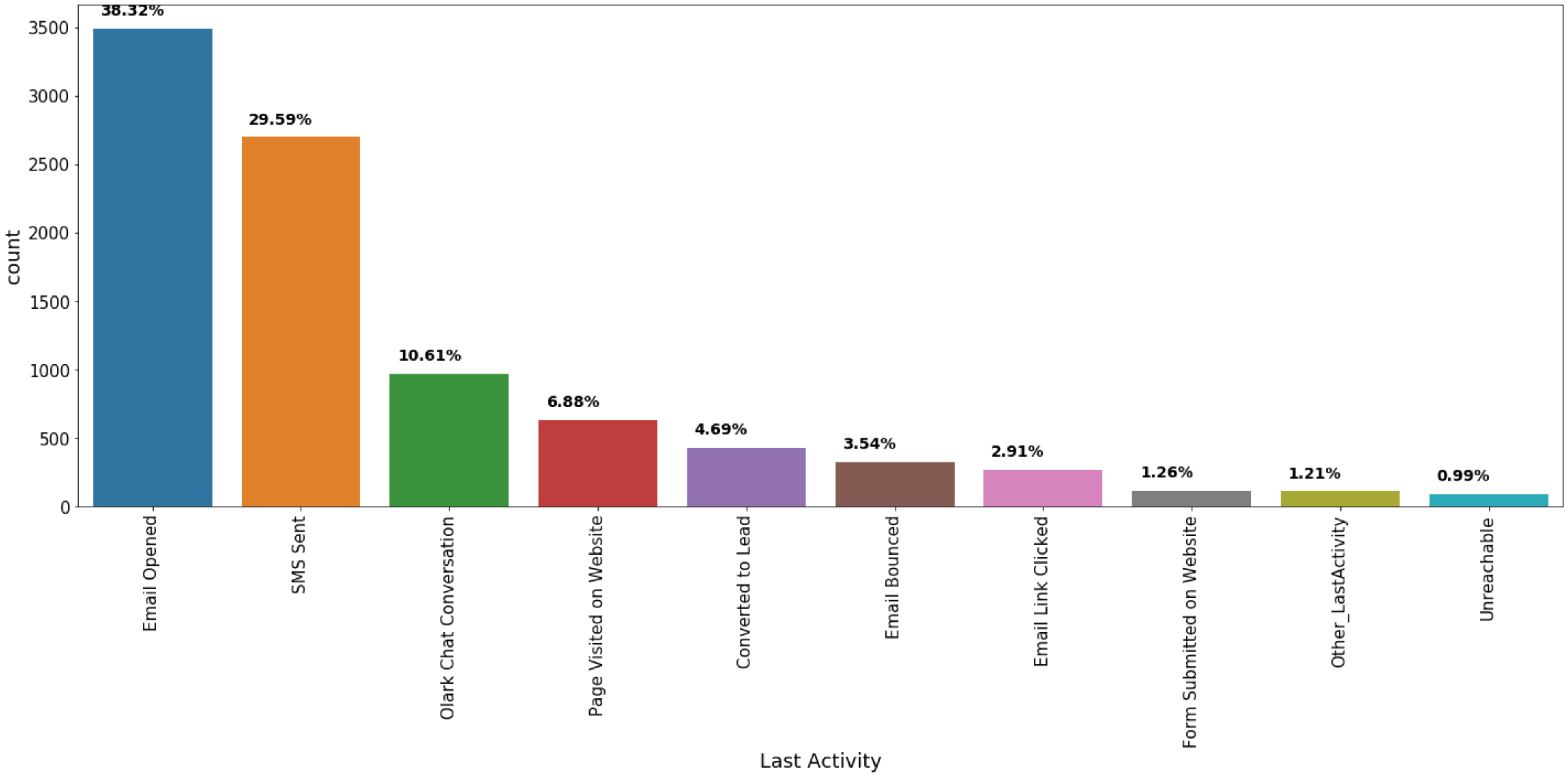
- Though 'Reference' & 'Welingkar Website' contribute just 5.86% and 1.56%, both of these have Maximum converison rate at Reference-91.76%, Welingkar Website-98.59%

Total time spent on website



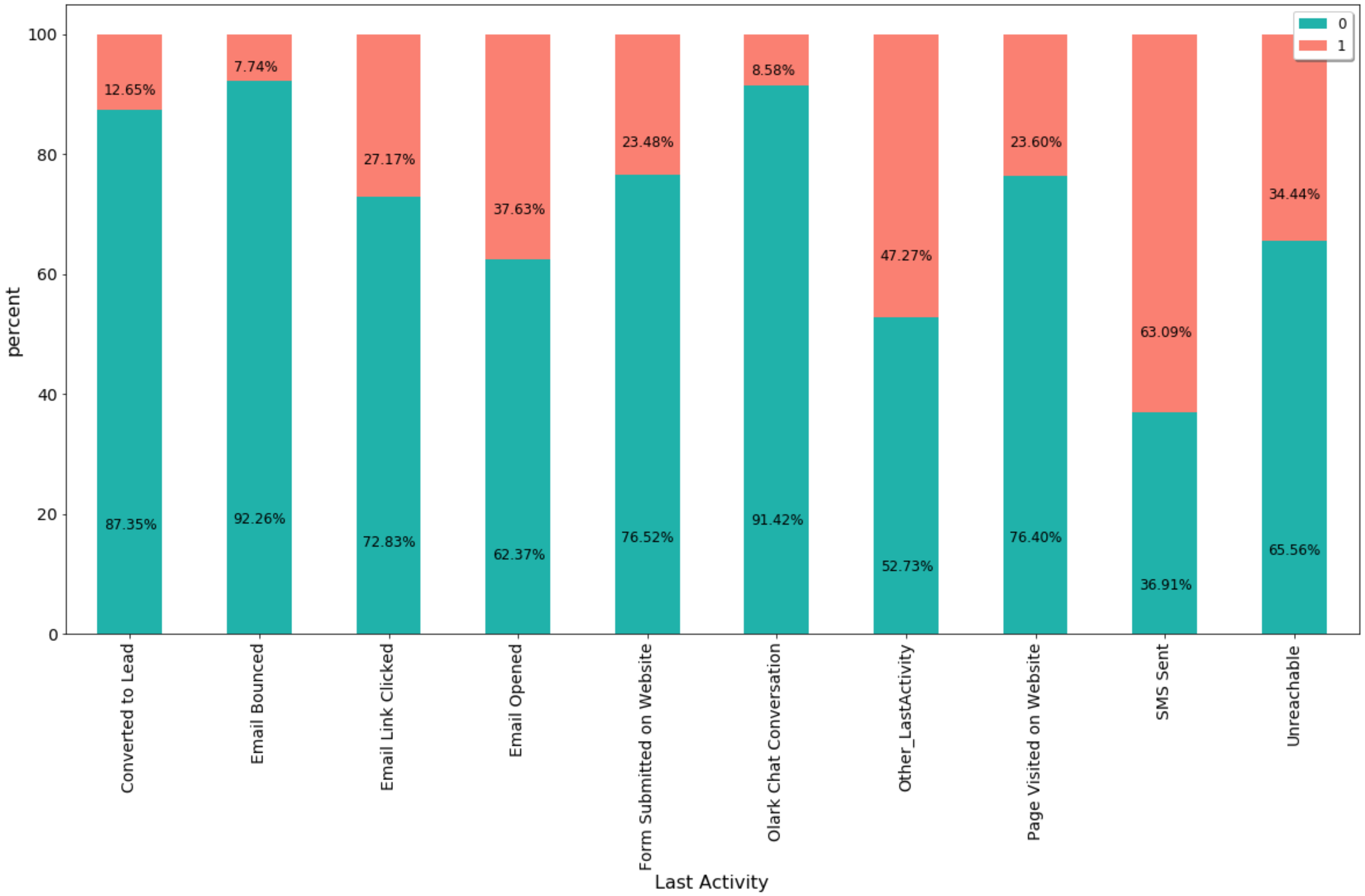
- For more Total time spent on website, Conversion is more.
- Hence people should be encouraged to spend more time on X education's website.

Last Activity



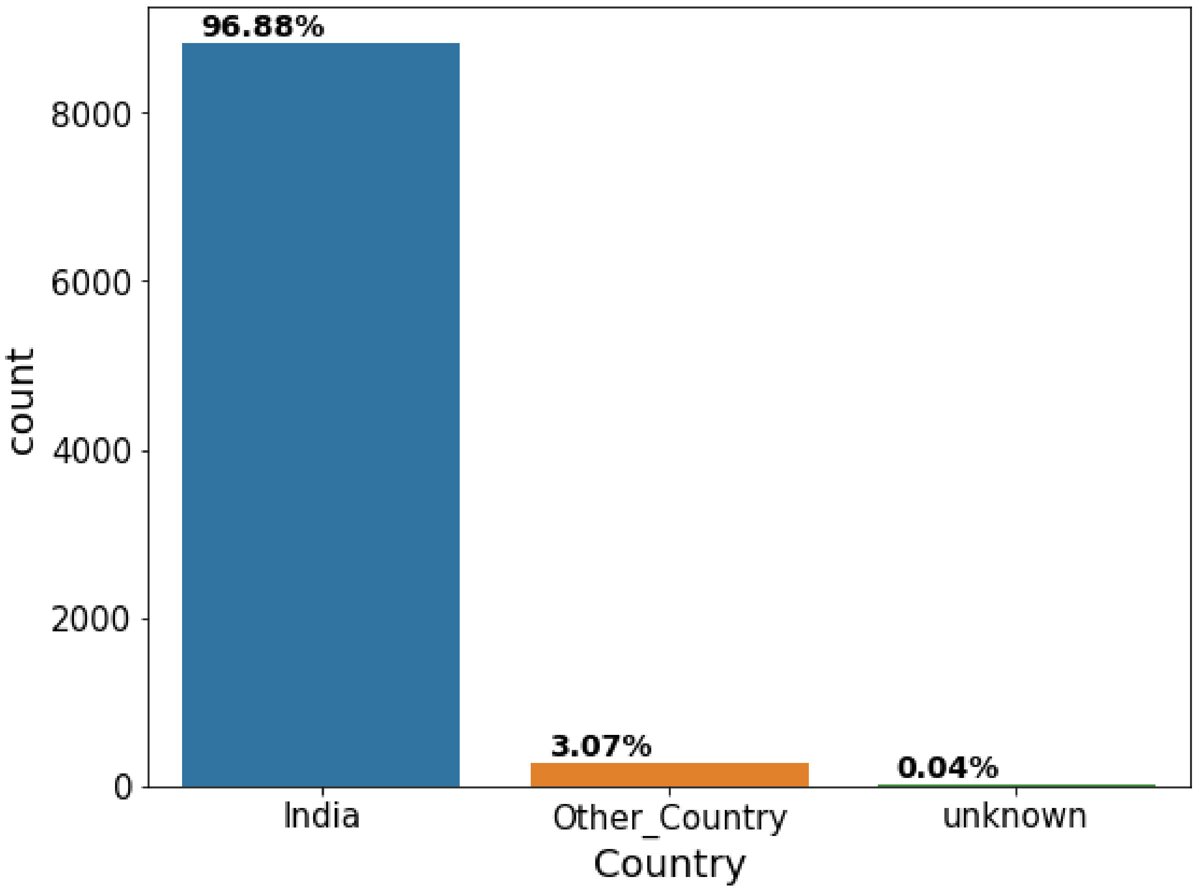
Order of Max to Min- Email Opened -> SMS Sent -> Olark Chat Conversation -> Page Visited On Website -> Converted to Lead -> Email Bounced -> Email Link Clicked -> Form Submitted On Website -> Other_LastActivity -> Unreachable

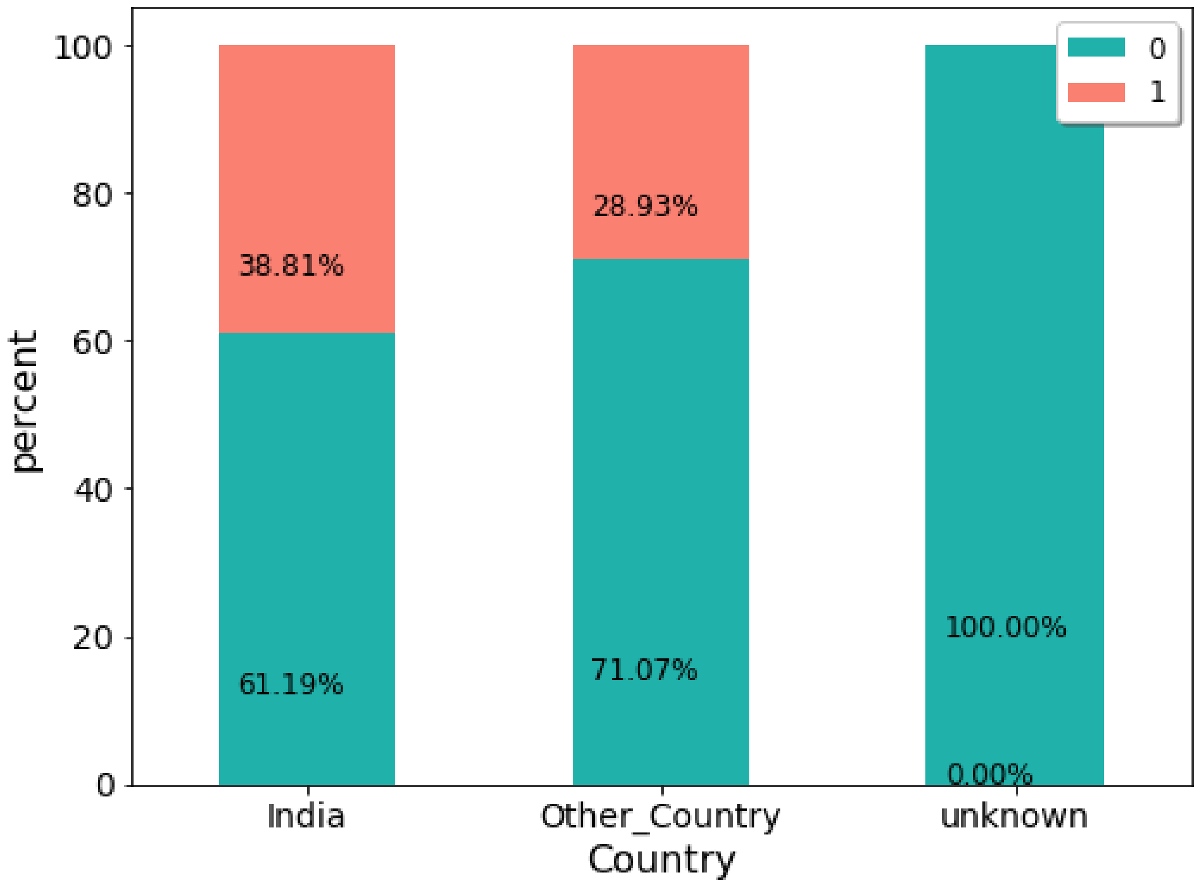
Last Activity



Maximum Conversion (63.09%) is for the people to whom Last Activity is SMS sent. This is followed by Other_LastActivity at 47.27%.

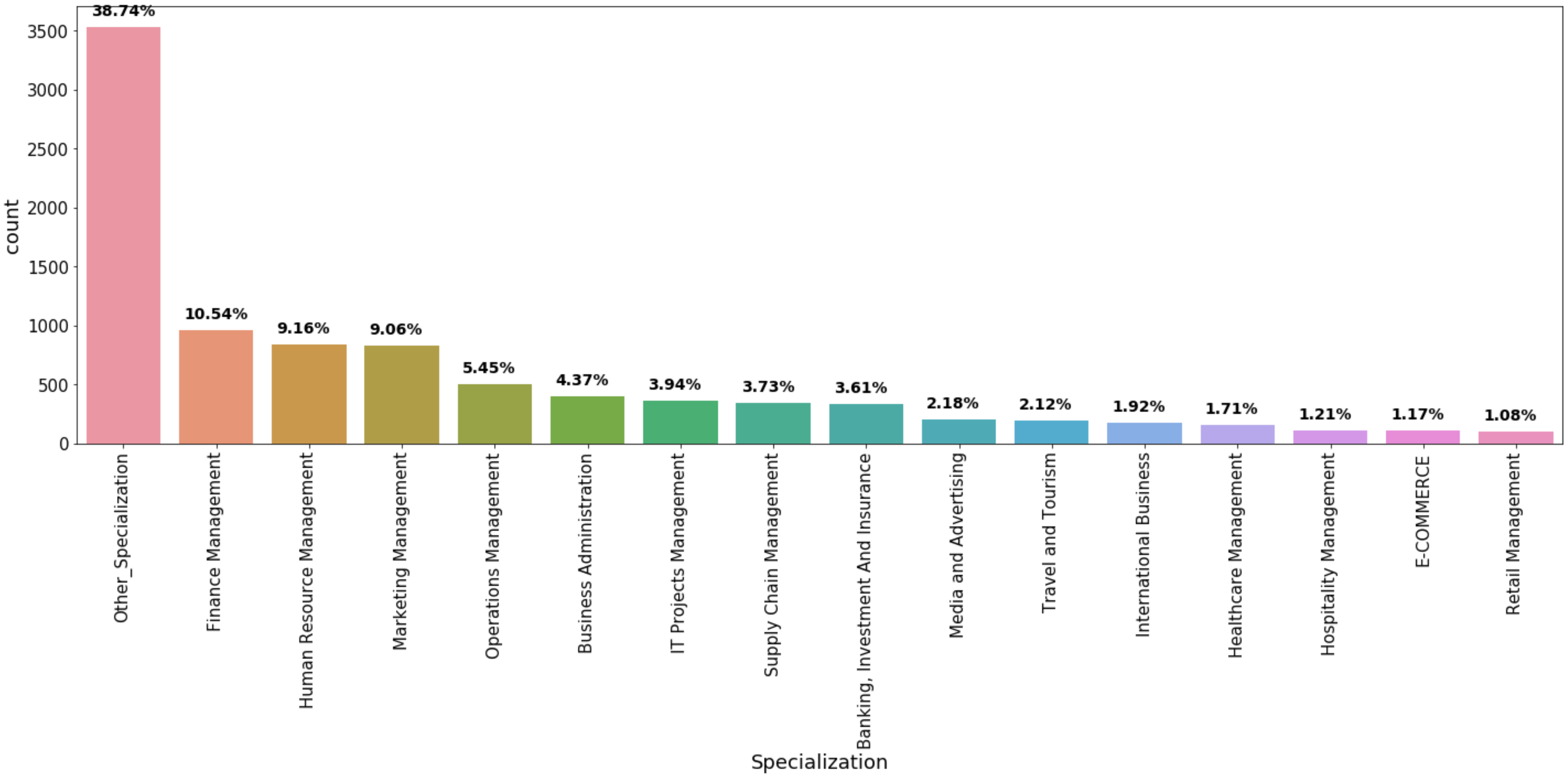
Country





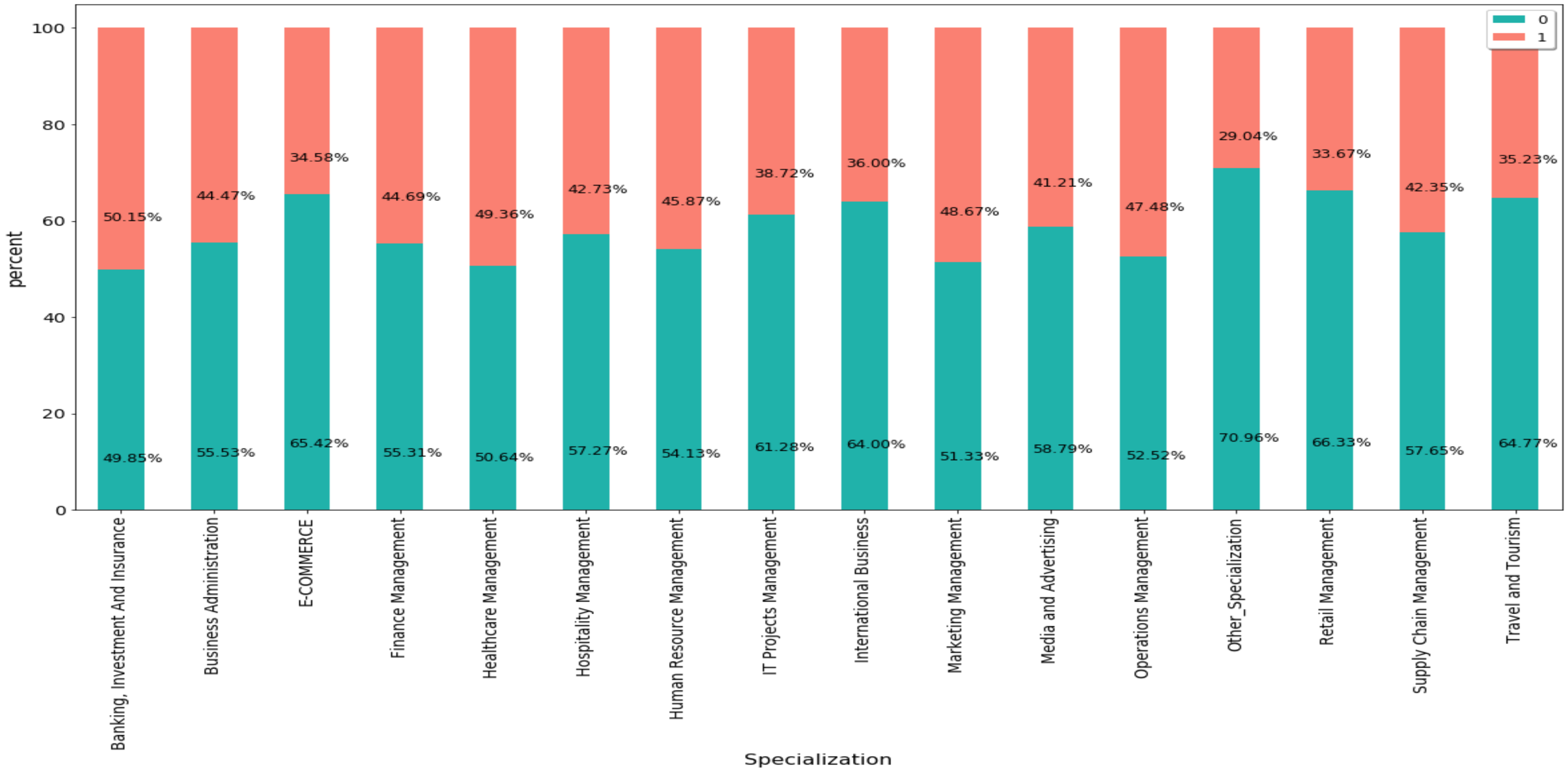
- Conversion rate for India is higher at 38.81% while for Other Countries it is 29%. Hence X education should concentrate better on Indian prospective leads.
- Prospects from unknown whereabouts have no conversion at all and X education can stop wasting resources on them.

Specialization



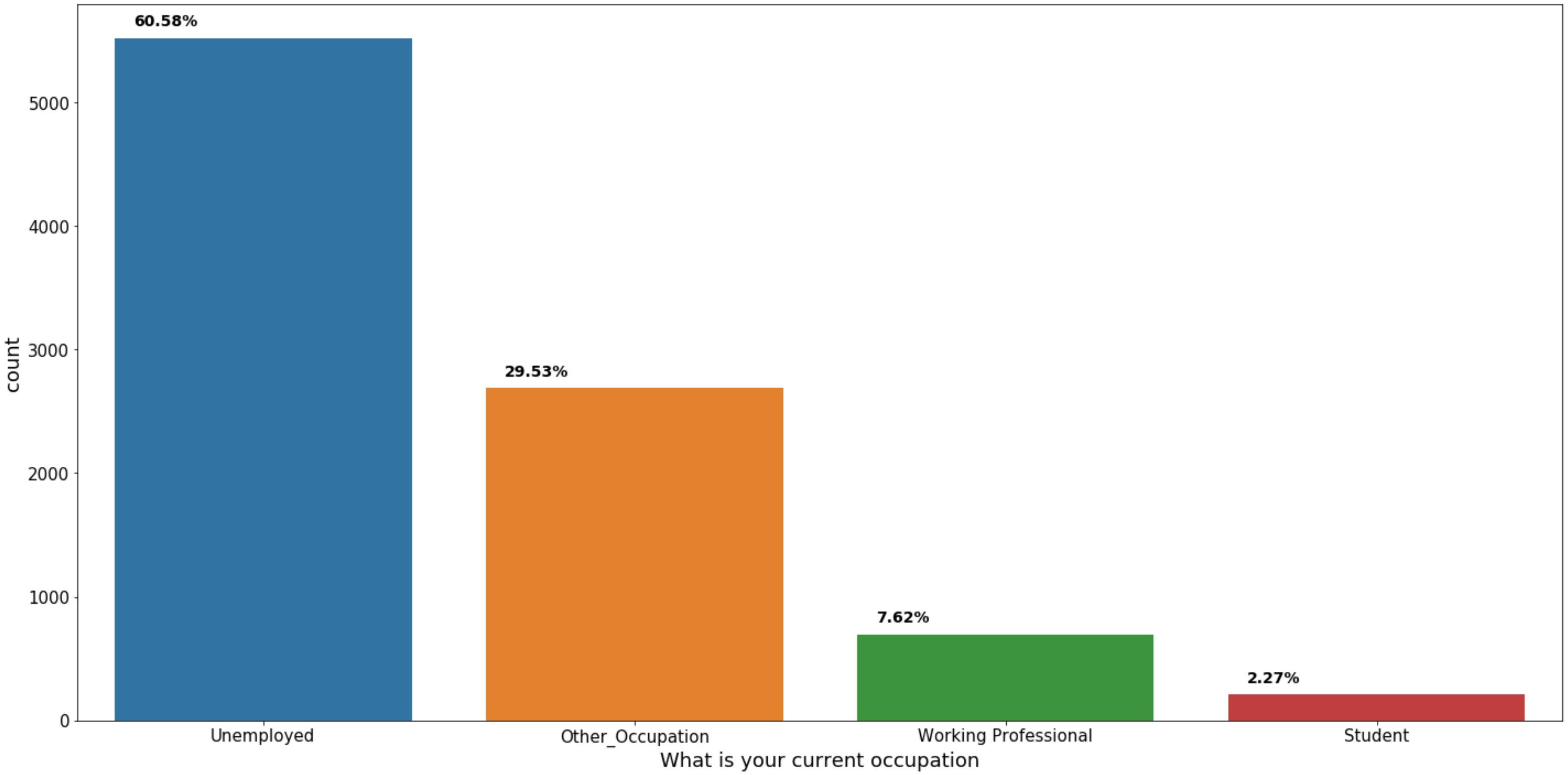
- Maximum percentage of data comes from category 'Other_Specialization' followed by Finance Management at 10.54%.
- All other categories are contributing less than 10%.

Specialization



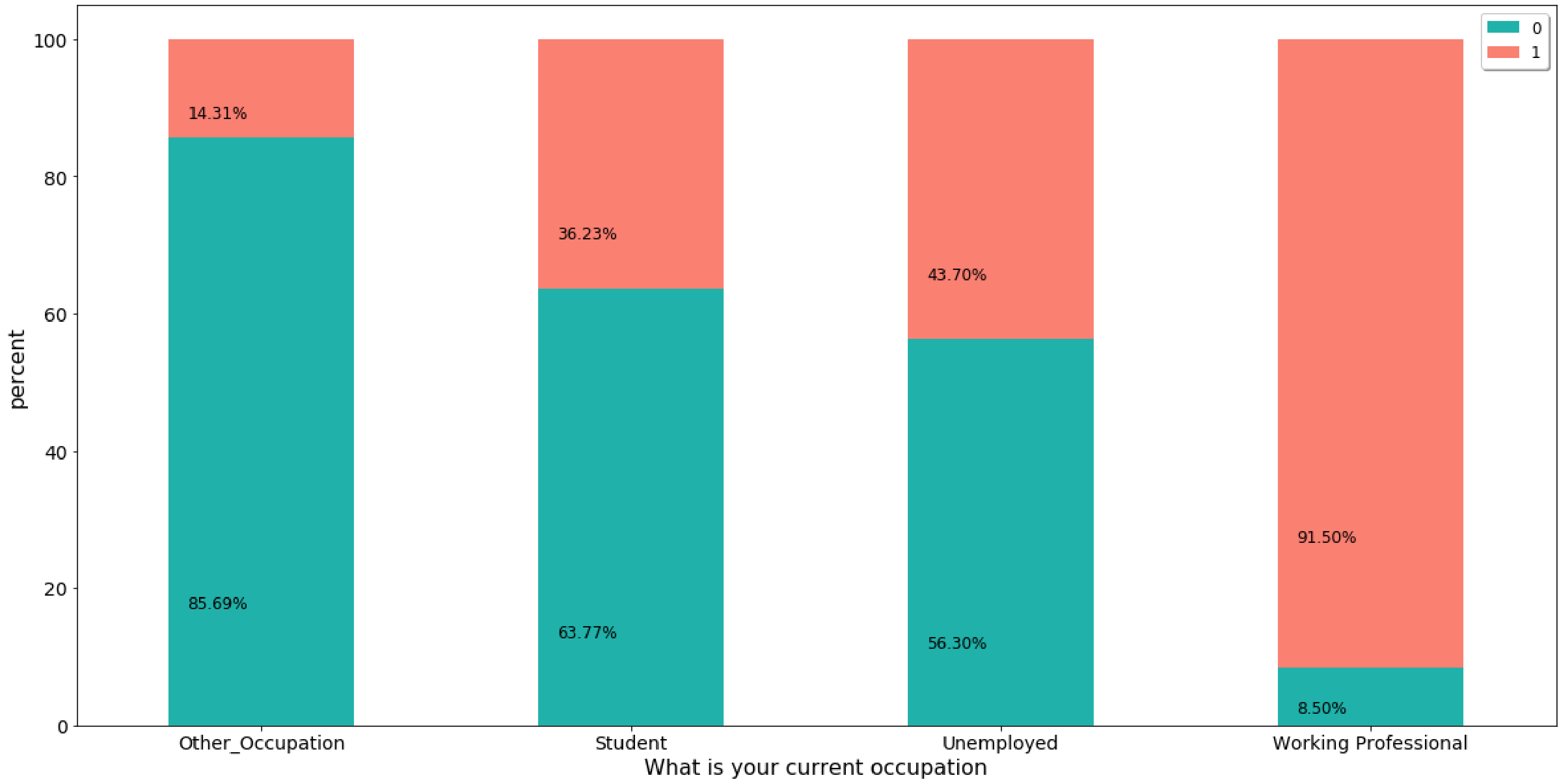
- Maximum Conversion is from Banking, Investment & Insurance at 50.15% followed by Healthcare Management at 49.36%.
- Marketing Management is close by at 48.67% and Operations Management at 47.48%.

What is your current occupation?



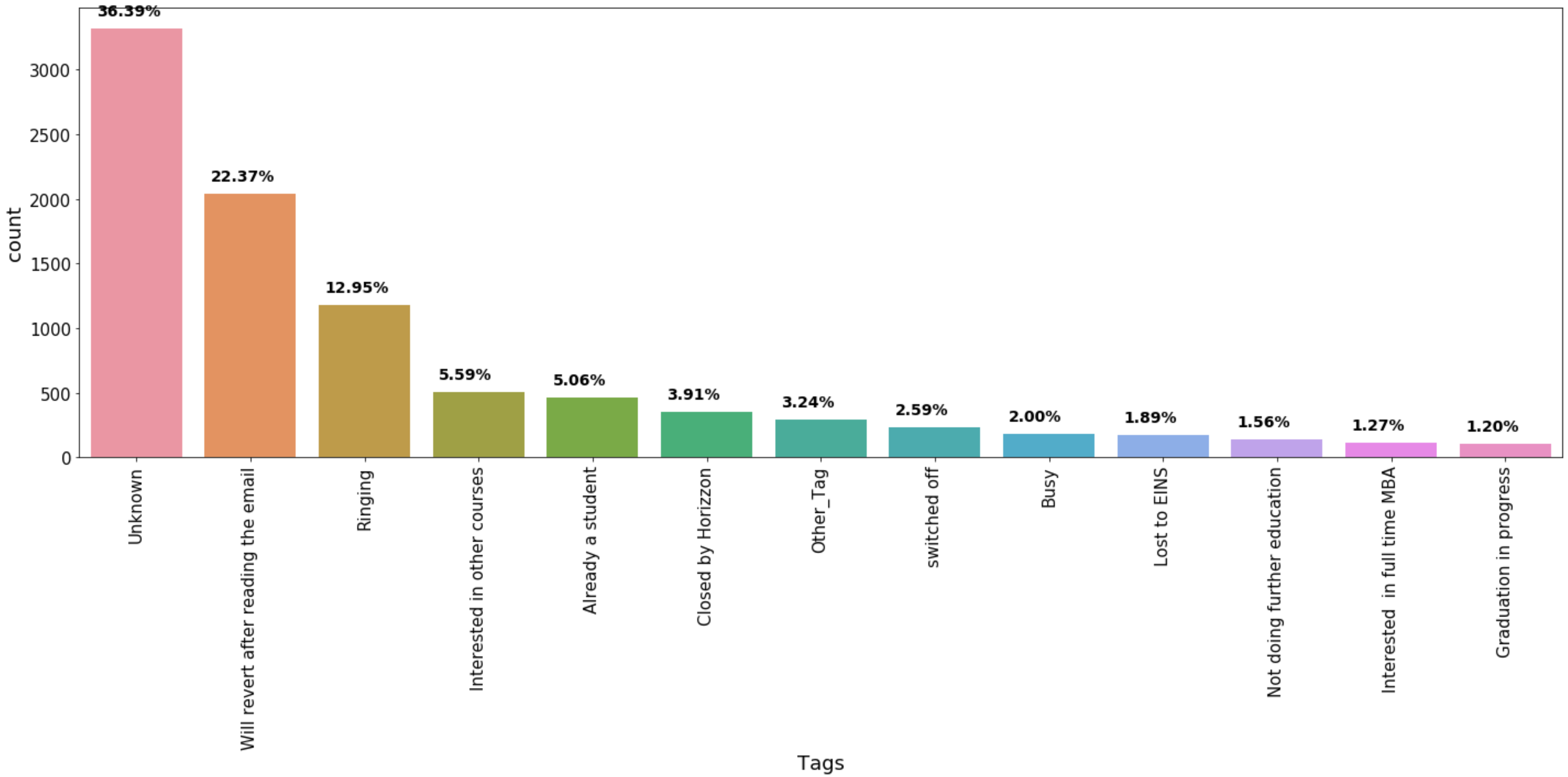
Maximum data i.e. 60.58% is from Unemployed people followed by Other_Occupation at 29.53%.

What is your current occupation?



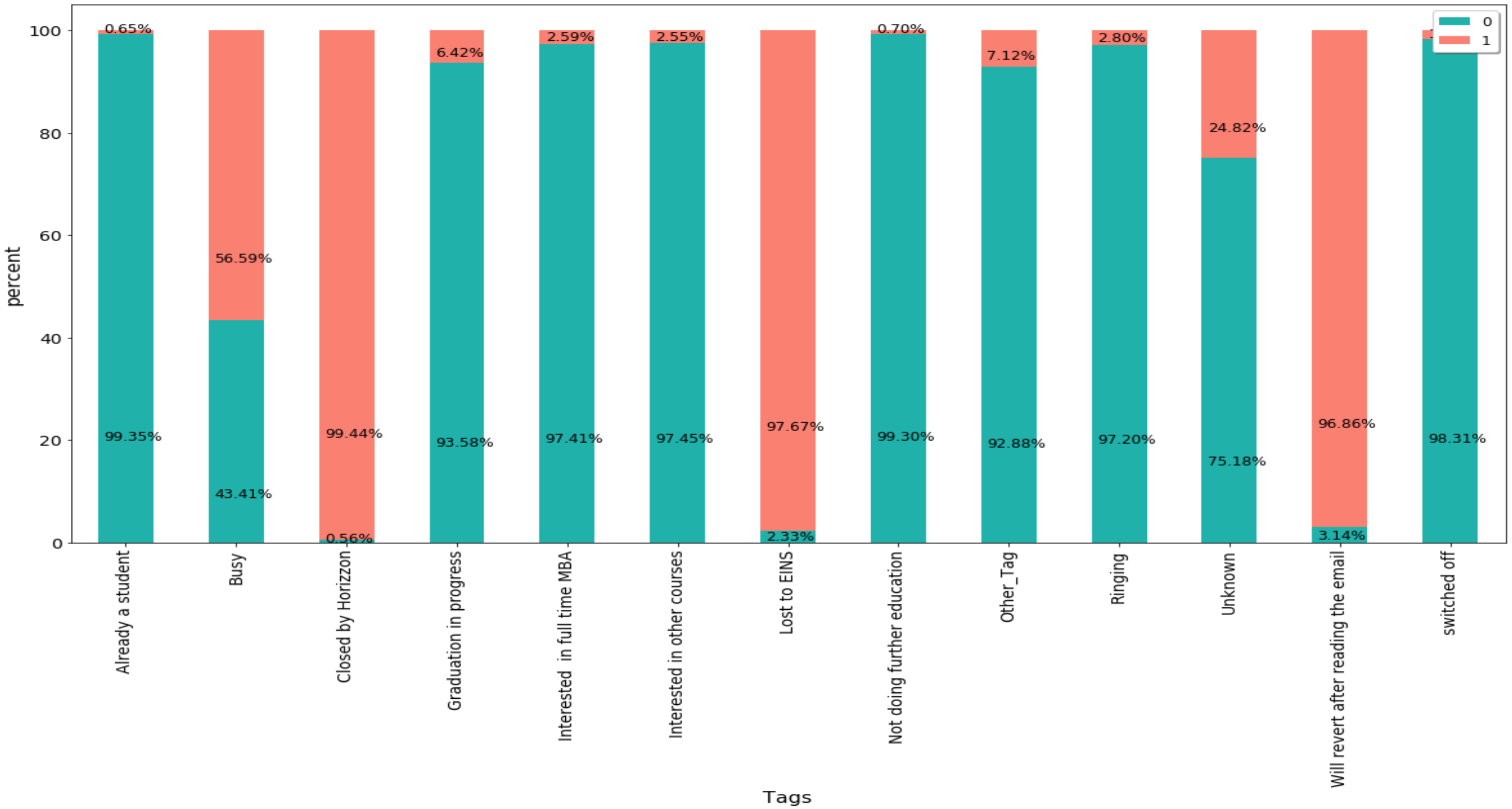
- The Conversion rate is huge at 91.50% for working Professional who represent just 7.62% data. Hence X Education needs to increase its reach to working professionals.
- Unemployed people have a modest conversion rate of 43.70% and X edu. might concentrate more in convinving them.
- Other_Occupation provides just 14.31% conversion and is a low focus area.

Tags



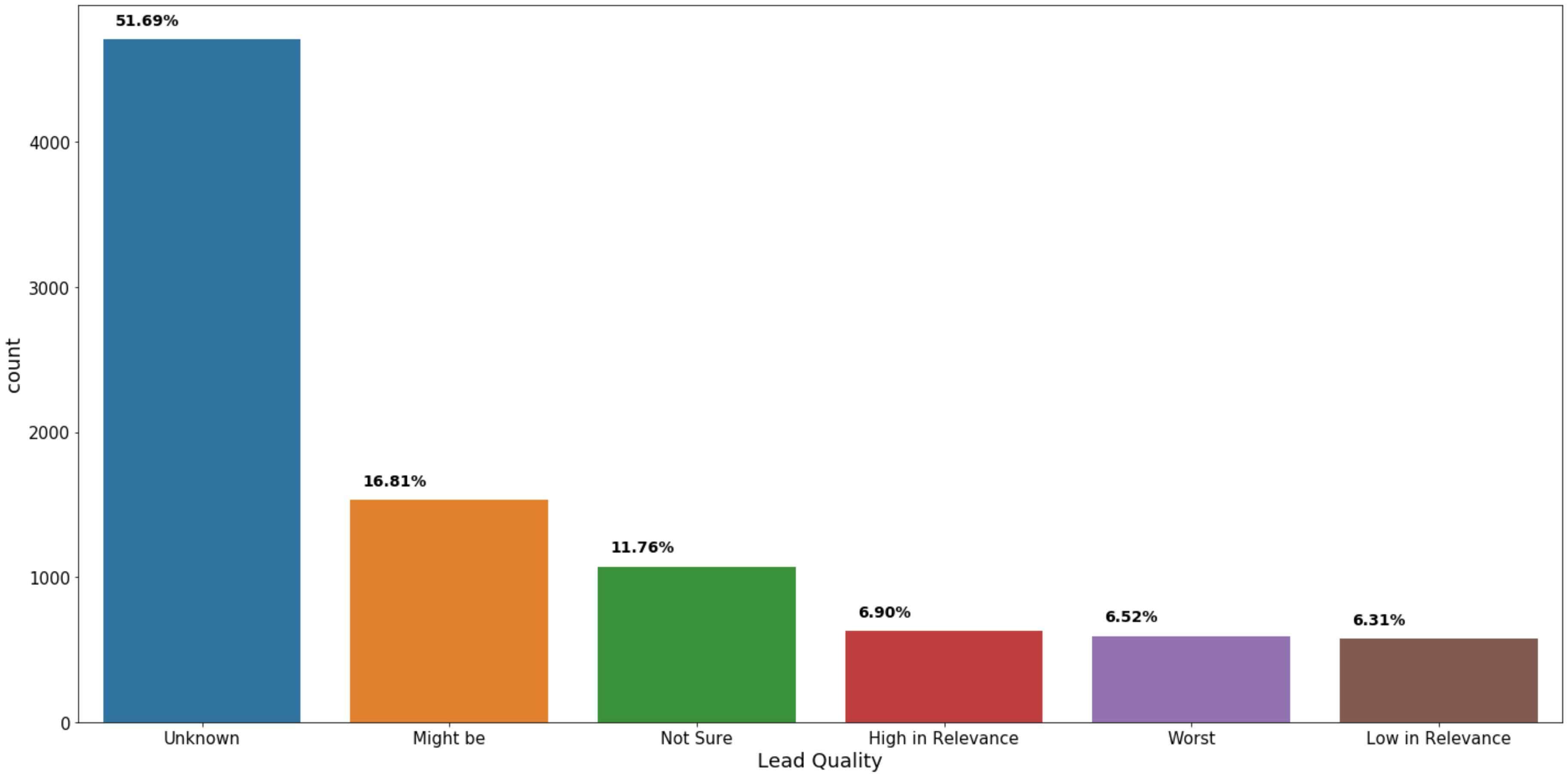
'Unknown' tags are maximum at 36.39% followed by 'will revert after reading the email' at 22.37%.

Tags



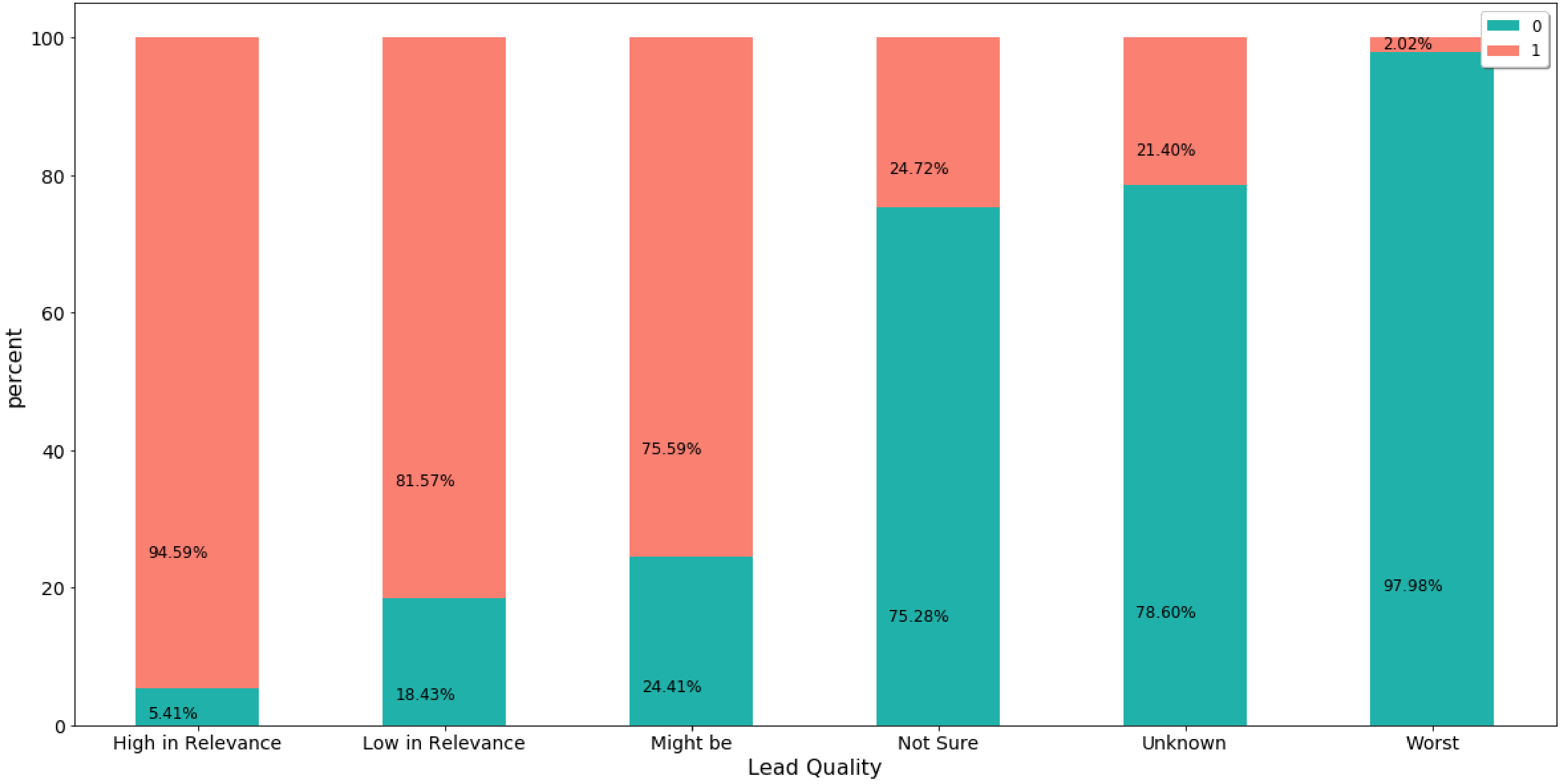
- Tags of 'Closed by Horizon' has a staggering conversion rate of 99.44% and this tag is 3.91% of total data.
- Second is 'Lost to EINS' at 97.67%,this tag represents 1.89% of toal data.
- X education needs to look into its Tags assignment methodology. The customers who were believed to have been shifted to other companies have actually converted with a huge percentage.

Lead Quality



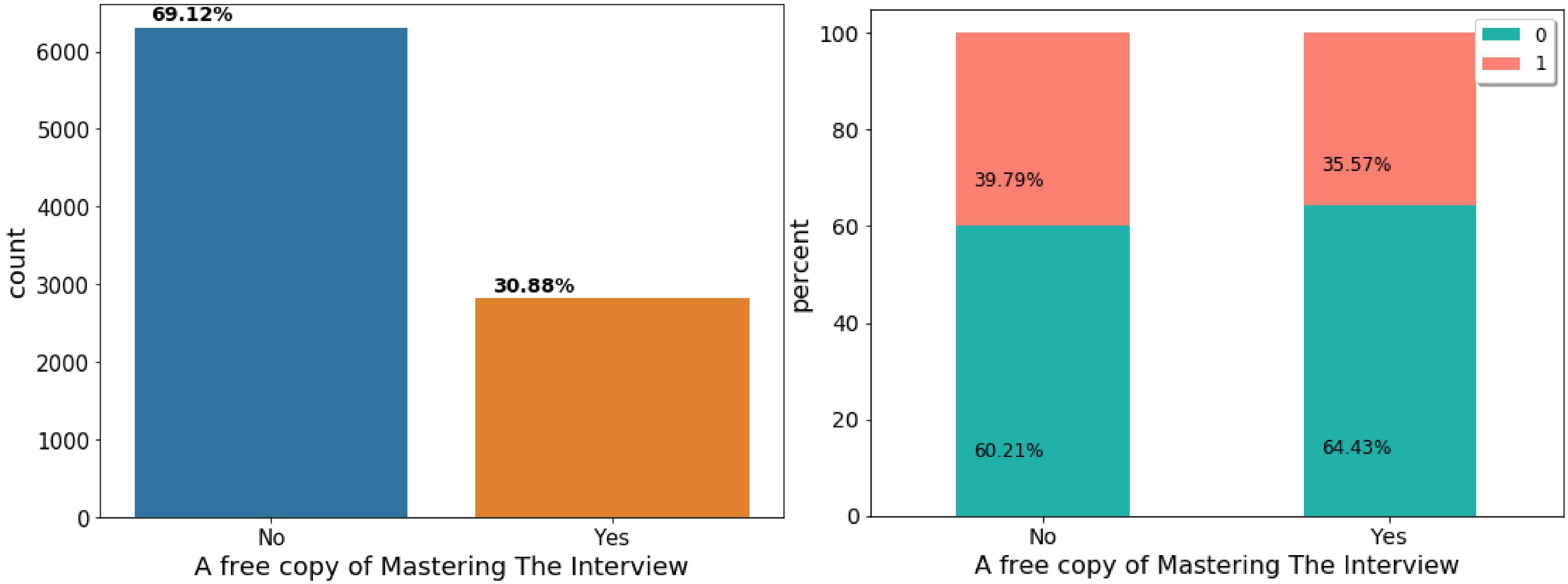
51.69% of data is classified as ‘Unsure’ followed by ‘Might be’ at 16.81%.

Lead Quality



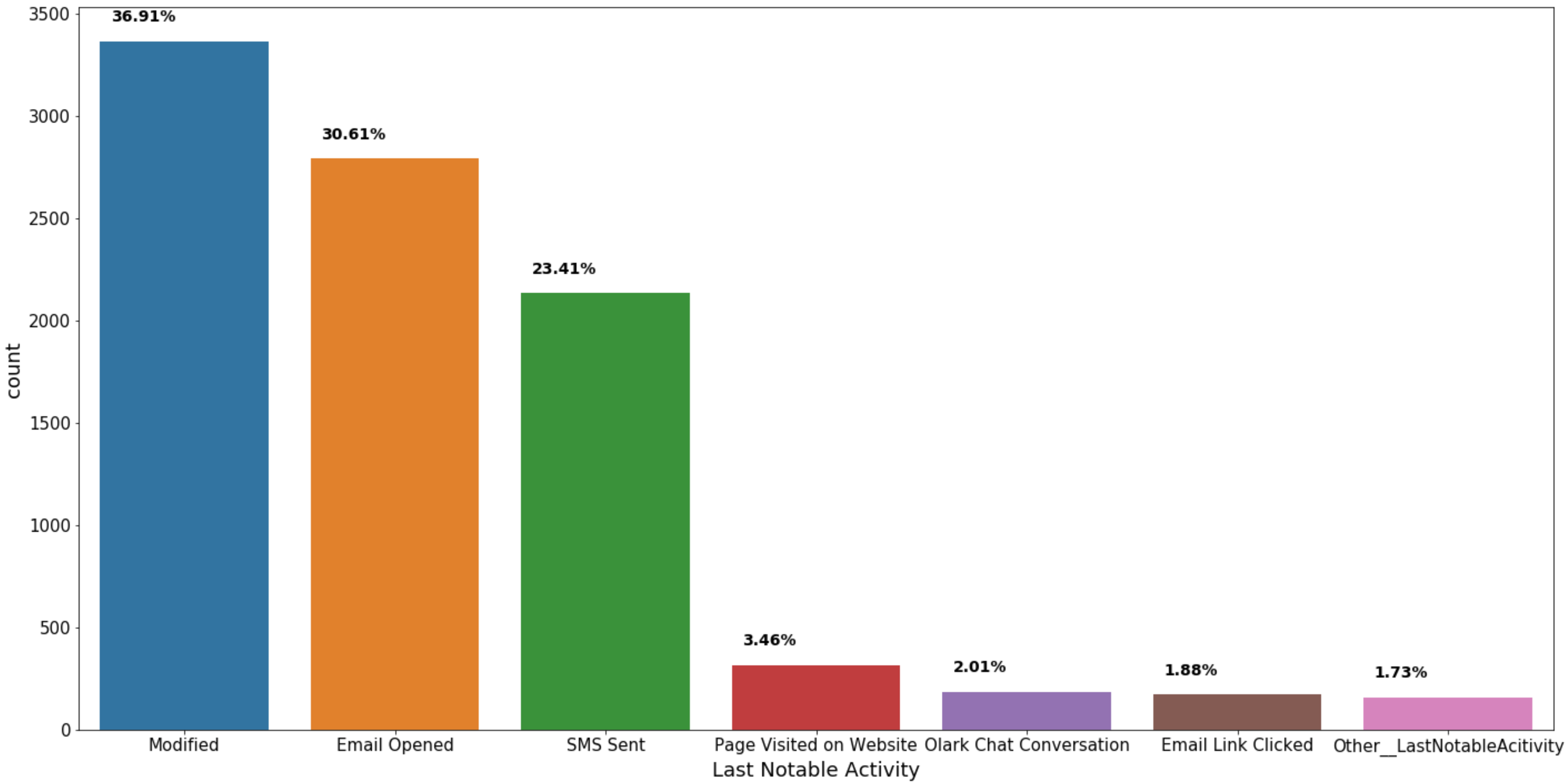
- The Lead Quality of 'High in Relevance' has huge conversion rate of 94.59% but represents just 6.9% of total data and needs to be paid more attention.
- The category maximum present i.e. Unknown has only 21.40% Conversion indicating a low potential area.

A free copy of mastering the interview



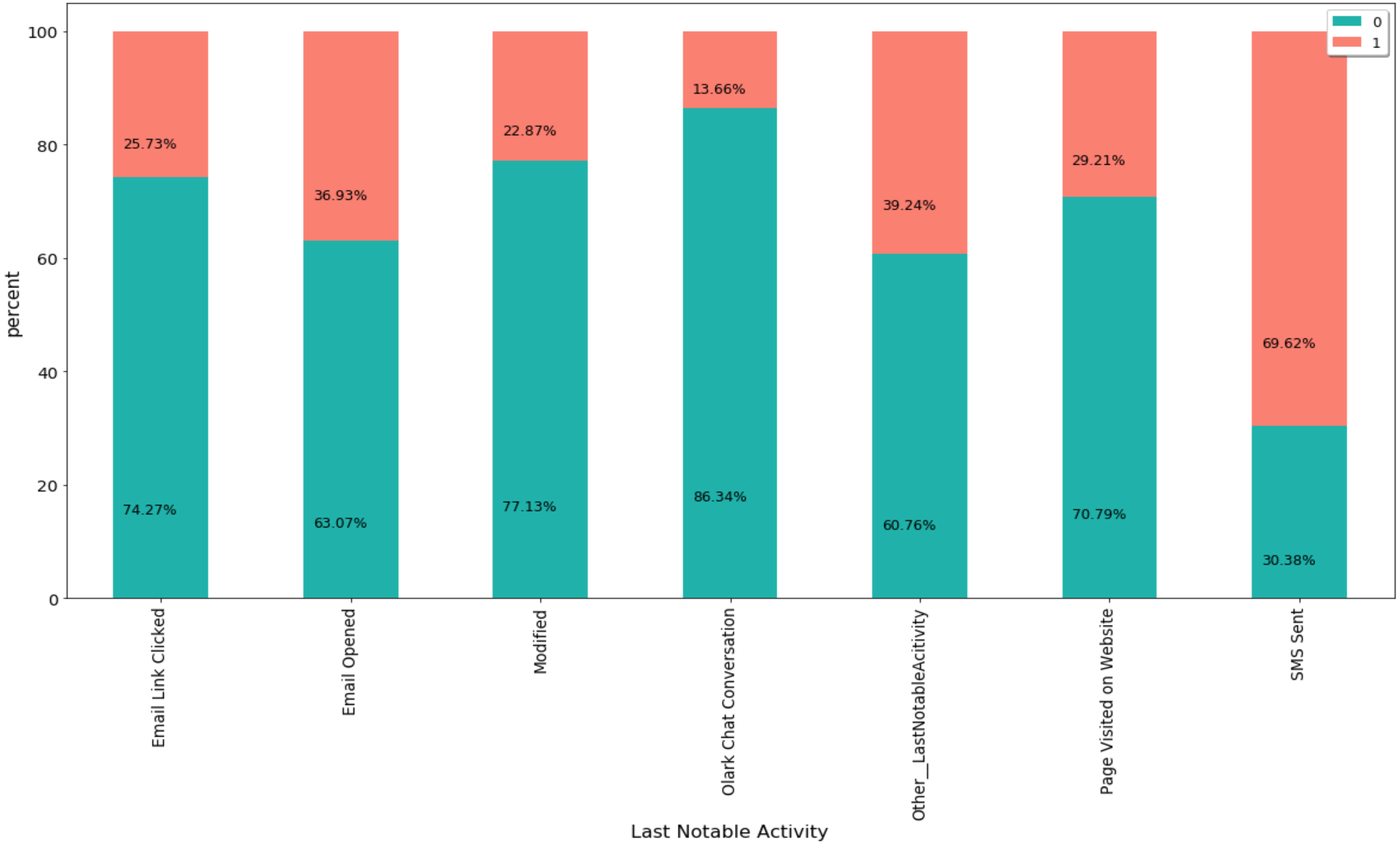
- This variable is not making any difference on conversion rate with customers who are not taking this having a higher conversion at 39.79%.
- Hence X education may stop offering this to customers in presence of any significant impact seen.

Last notable activity



36.91% of data is Modified Last Notable Activity, followed by Email opened at 30.61%

Last notable activity



- Conversion rate is highest for SMS sent at 69.62% which is 23.41% of total data.
- Maximum present category Modified has just 22.87% . Hence X education should send more SMS to leads. And consider Modified as low potential area.

Dummy variables & Numerical encoding
Train-Test split

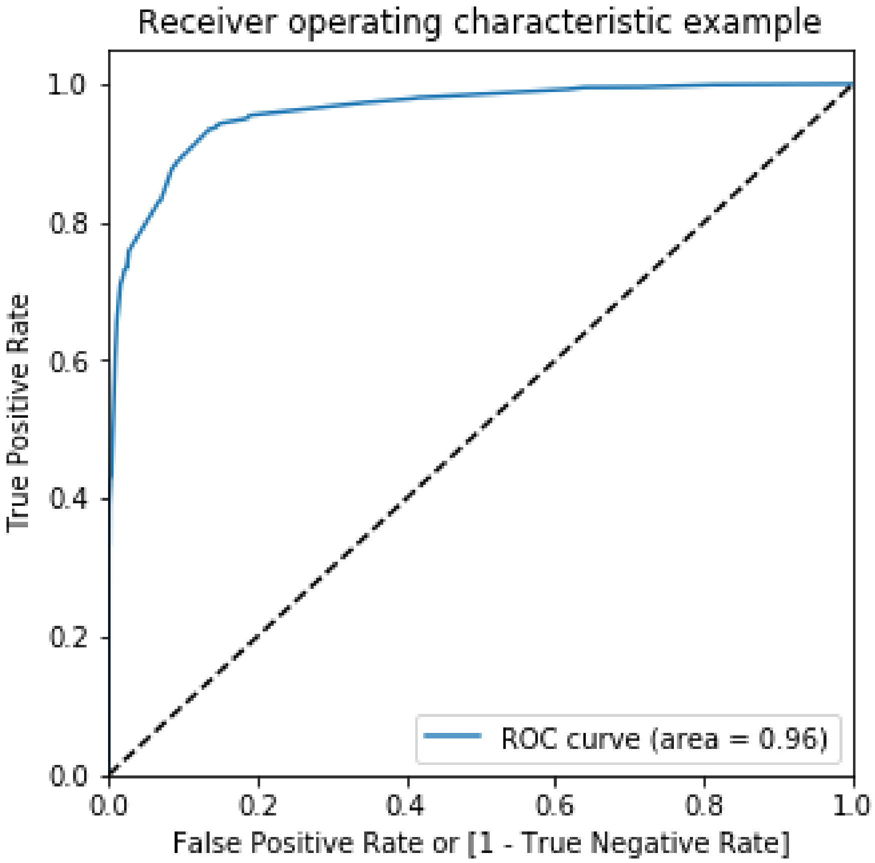
- To proceed with Logistic regression, Dummy variables are created.
- Original Categorical variables are dropped after dummy creation.
- Yes & No values in columns are converted to 1 & 0 respectively.
- The final dataset contains-
- Rows: 9112
- Column: 68
- Final dataset is split into train and test dataset in 70%-30% proportion.
- Train & Test data are split into X and y.
- y is taken as ‘Converted’, remaining variables as X.

Model Building

- 20 Features were selected using RFE.
- Five Logistic regression models were built iteratively
- Final model was selected based on
- p-values <0.05 for all variables, indicating significance
- VIF < 5, indicating absence of multicollinearity
- Model performance measures-
- High values of Accuracy, Sensitivity & Specificity indicate good predictive powers of model.
- Low False positive rate indicates model’s ability to predict Positive values accurately.

Accuracy	89.19%
Sensitivity	83.07%
Specificity	93.08%
False Positive Rate	6.91%
Positive Predictive Value	88.39%
Negative Predictive Value	89.66%

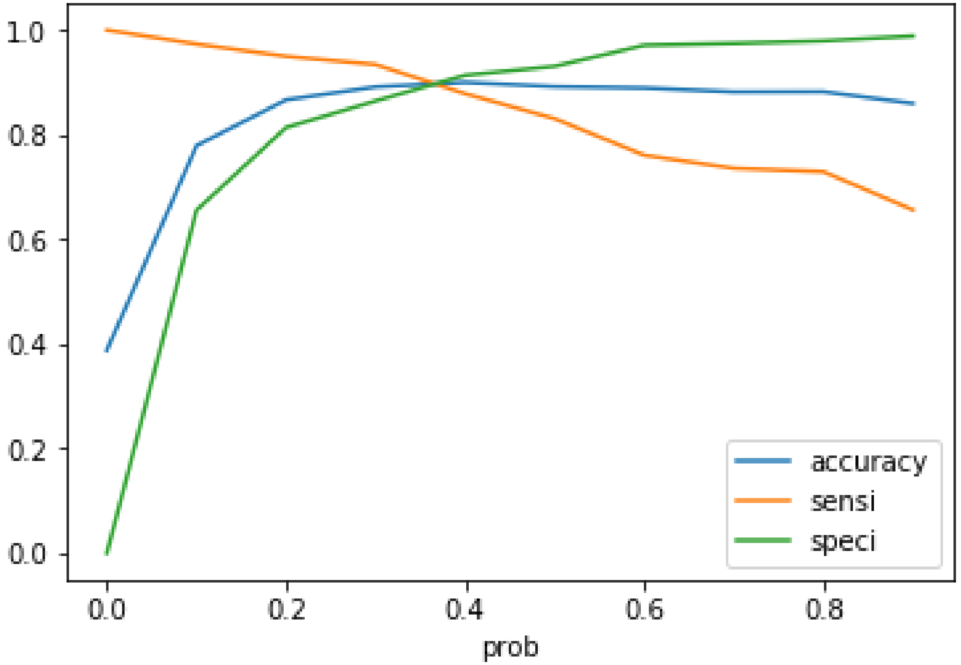
ROC Curve



- The ROC curve is towards the upper-left corner, with a high area under curve.
- This indicates good predictive powers of model.

Model accuracy check

- Accuracy, Sensitivity & Specificity plot to find optimum cutoff for probability



Accuracy	89.99%
Sensitivity	88.68%
Specificity	90.83%
Positive Predictive Value	85.97%
Negative Predictive Value	92.68%
Precision	88.39%
Recall	83.07%

- The three curves intersect at ~0.38.
- Model accuracy at this point is 89.99%, which is very close to earlier calculated value.

Model fit on test data

- Final model was fit on the test data.
- Predictions of Converted values were made.
- The accuracy achieved on test dataset is also same at 90.16%.
- Sensitivity of 87.5% and Specificity of 91.77% was achieved.
- These measures indicate a good fit of model on the test data as well.

Conversion

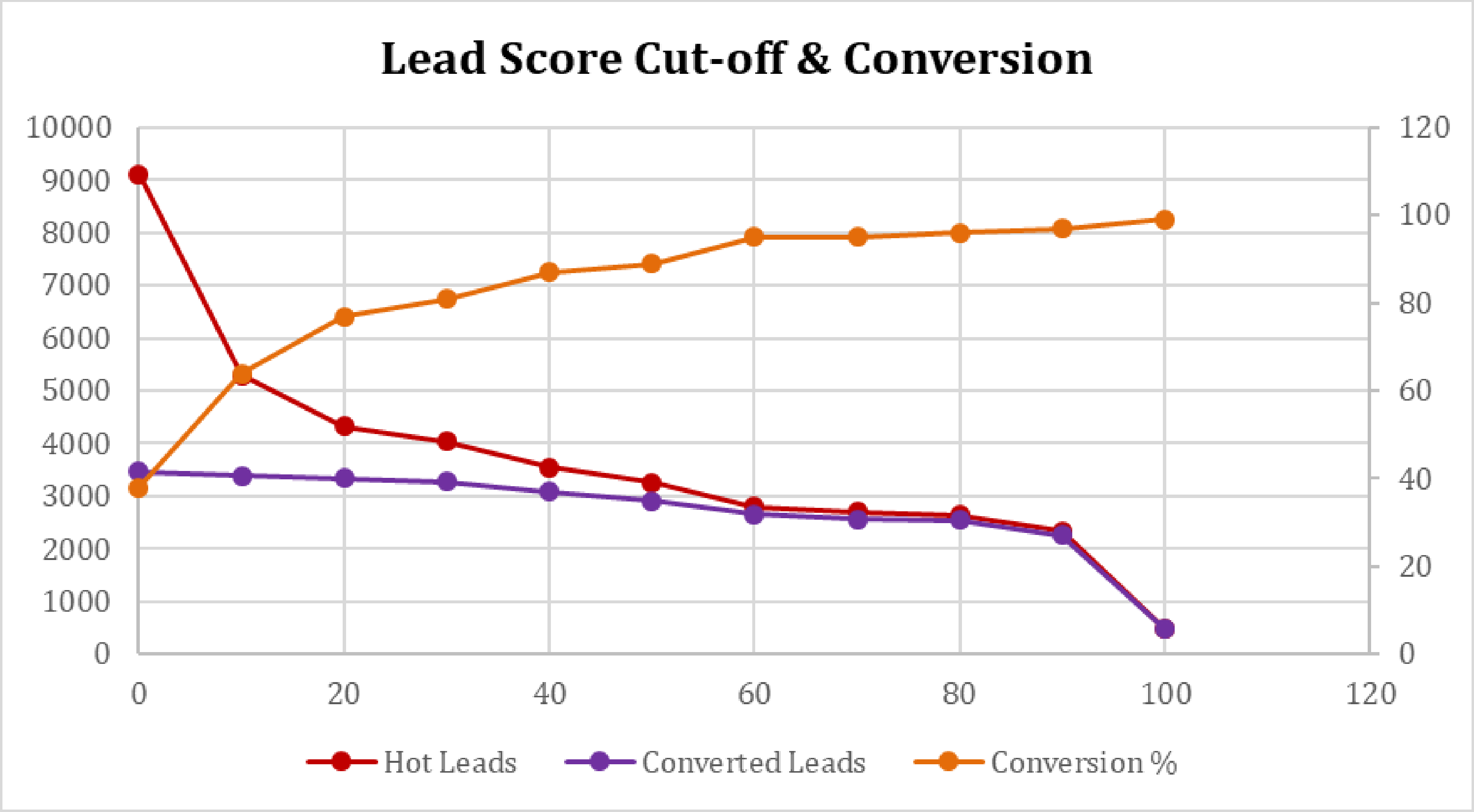
- To calculate Conversion on the entire dataset, a master data frame was created with final y(s) from train and test sets.
- From train, ‘y_train_pred_final’ and from test, ‘y_pred_final’ are concatenated
- Cutoff Lead Score was applied on this dataset to select only Hot leads
- At Lead Score of 30, Conversion of 81% was achieved, which is more than target of 80%
- Conversion % were checked at different Lead Score cut-offs & tabulated next-

Conversion rates at different Lead score cut-offs

Lead Score Cut-off	Hot Leads (Number)	Actual Converted Leads (Number)	Conversion (%) Actual
0	9112	3463	38
10	5295	3389	64
20	4331	3335	77
30	4043	3275	81
40	3539	3079	87
50	3262	2903	89
60	2801	2661	95
70	2698	2563	95
80	2643	2537	96
90	2332	2262	97
100	476	471	99

To achieve target conversion of minimum 80%

Conclusion



As the Lead score cut-off increases, conversion % increases.

Recommendations

- To get as many customers as possible in cases of additional man-power availability, X education must keep the lead score lower, starting at ‘0’. Whereas to achieve target conversion of grater than 80%, it should keep the cut off as 30.

- Thus, in the model, data frame ‘df’ in the end can be tweaked for cut off Lead Score to gauge the Conversion percentages w.r.t. actual converted.

- Lowering the lead score cut off reduces conversion %, but it increases number of actual converted.

- Based on the man power availability with X education, it may decide to give weightage to conversion % or actual numbers.

- To understand effect of individual variable’s classes, plots from EDA may be referred.

Recommendations

- Based on the model, the variables positively contributing to conversion are-

Tag	Coefficient
Tags_Lost to EINS	6.40
What is your current occupation_Working Professional	3.92
Lead Source_Welingak Website	3.32
Tags_Will revert after reading the email	2.71
What is your current occupation_Unemployed	2.43
What is your current occupation_Student	2.07
Last Activity_SMS Sent	1.97

- X education should focus on these six variables for the maximum conversion rates.