Lead Scoring Case Study

Submitted by:

Rakesh Khanna
Swati Kumar
Swati Patil
Batch July 2022

Problem statement

- An education company named X Education sells online courses to industry professionals. Their conversion rate is approx 38% as per the data.
- To improve the conversion rate, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future.

Expected outcome :

- 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.
- The model to accommodate few more problems presented by the company.

Approach

Business understanding :

The information of data from the ‘Leads data dictionary’ file is studied to get an overall picture of attributes.

Data cleaning :

- The data is checked for it’s shape, size, data type, missing value, formatting errors and data imbalance.
- Columns with missing data > 45% are excluded from the analysis.
- Imputation technique is suggested for categorical columns to replace ‘select’ with ‘null value’, missing country/city with mode value.
- As the last step, a small percentage of null rows was left. This was deleted.
- Outliers are identified and highlighted.

EDA :

- Univariable data analysis : Data distribution is checked.
- Univariable data analysis : Correlations are checked.

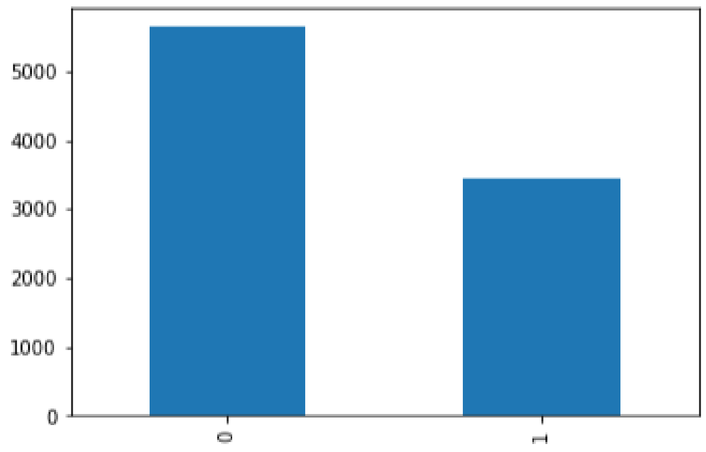
Model building :

- Logistic regression method is used for model building and prediction.

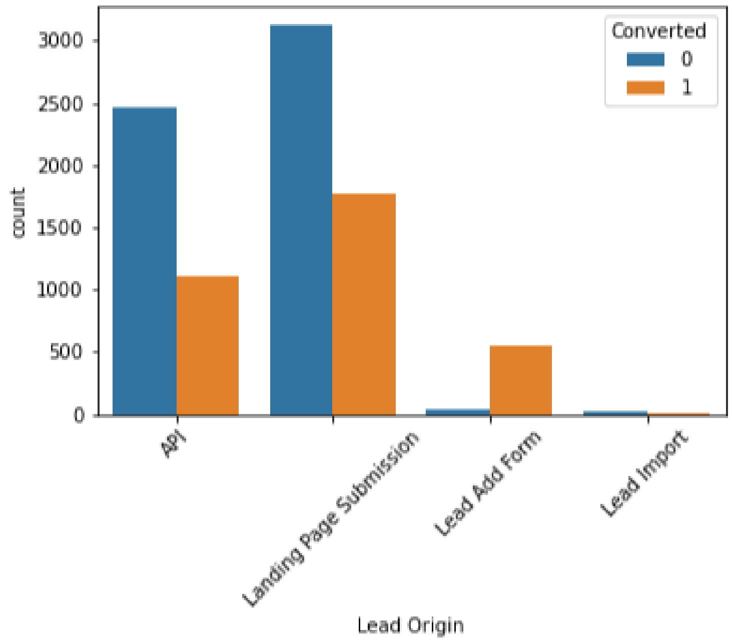
Model evaluation

Prediction

Data analysis :

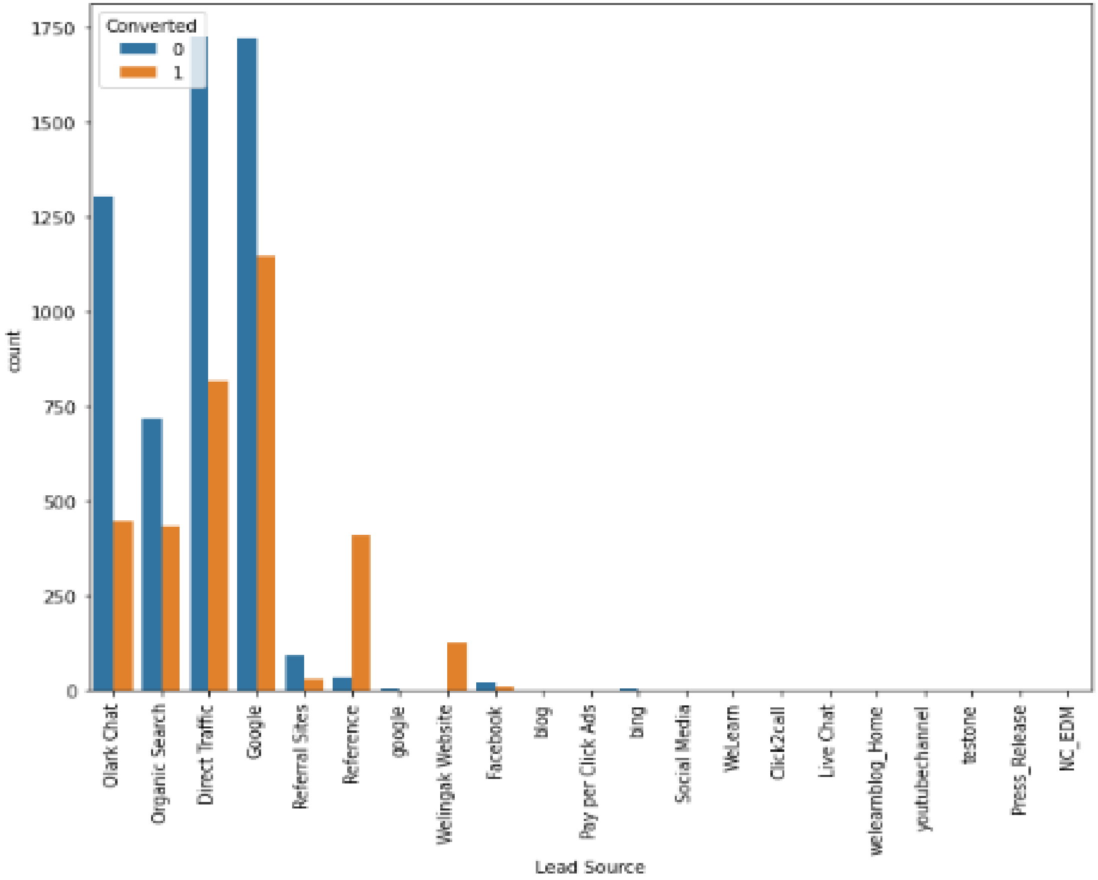


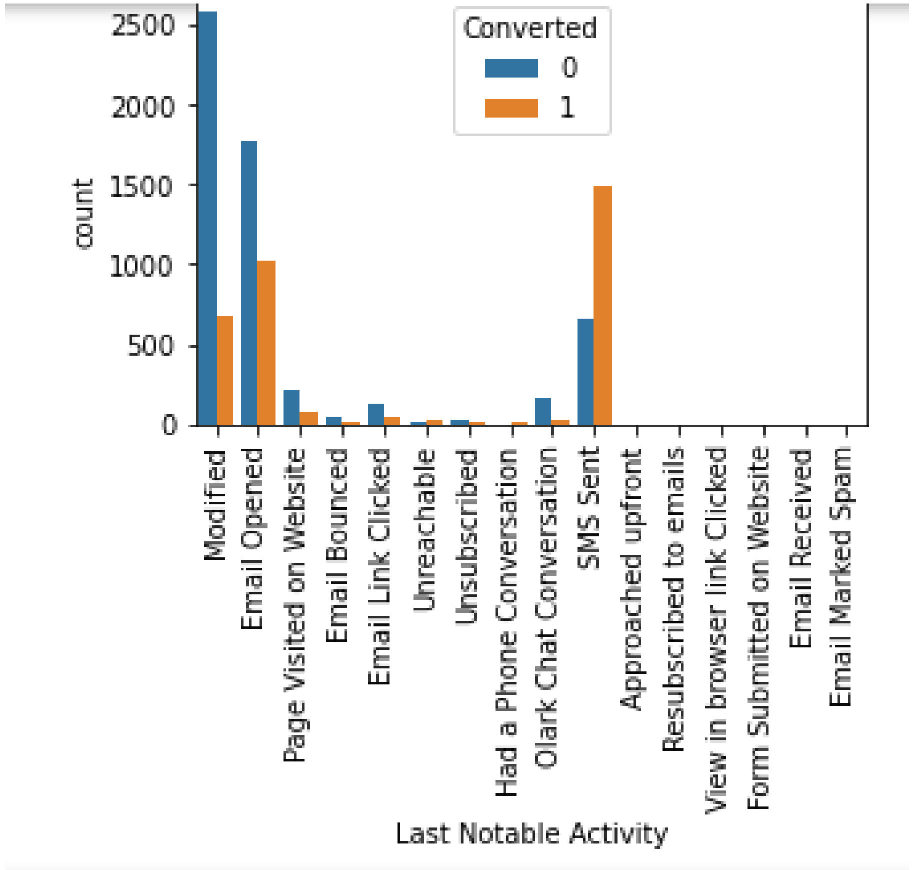
Approximately 38% of leads are converted.



The maximum conversion is done through 'Landing Page Submission' followed by 'API' and 'Lead Add Form'.

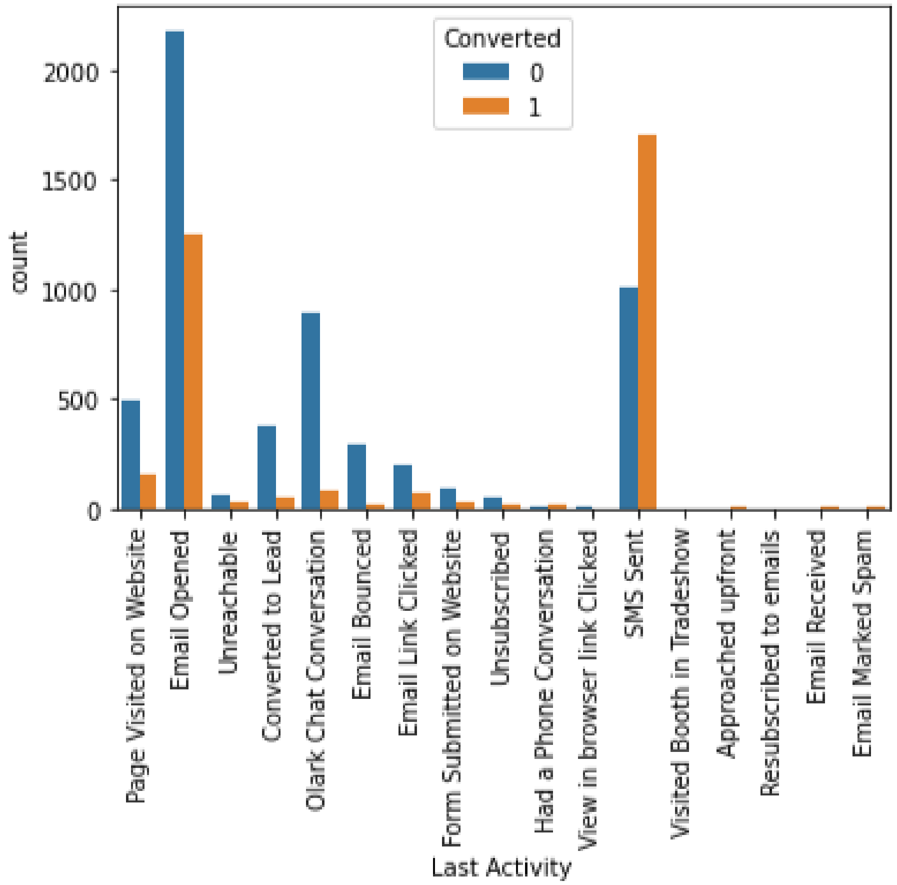
Data analysis :



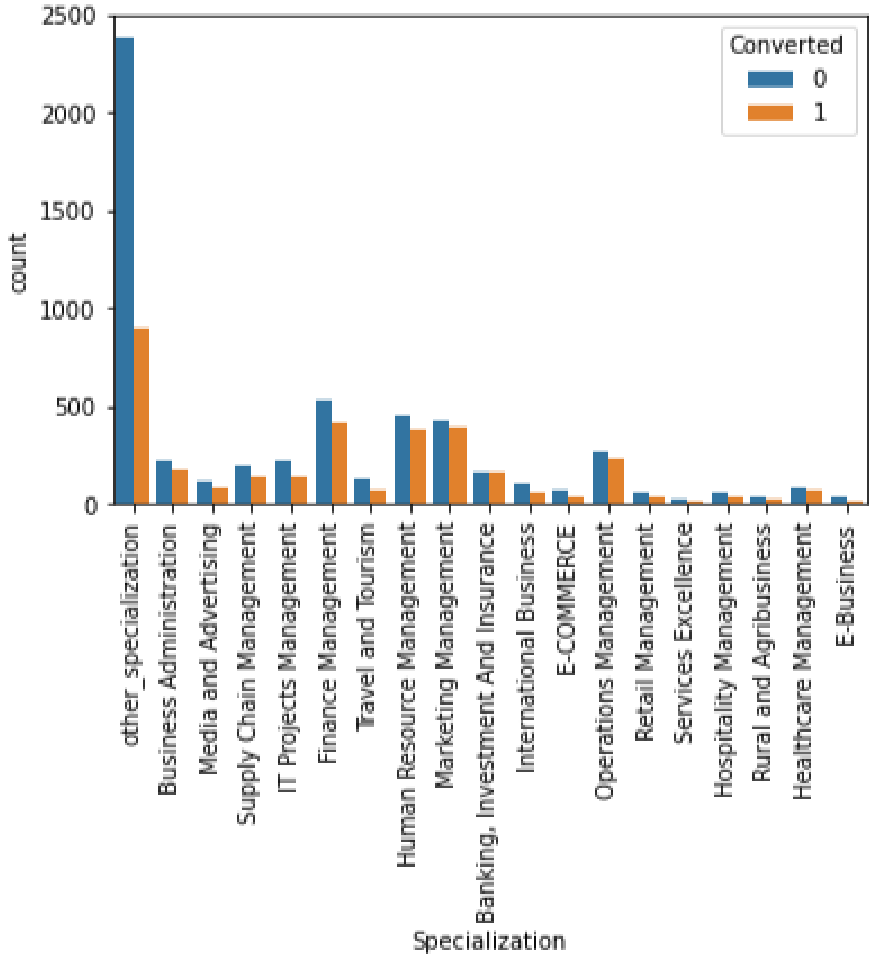


The maximum conversion is done through 'Google', 'Direct traffic'.
Conversion rate from 'Reference' and 'Welingak Website' are higher.

Data analysis :

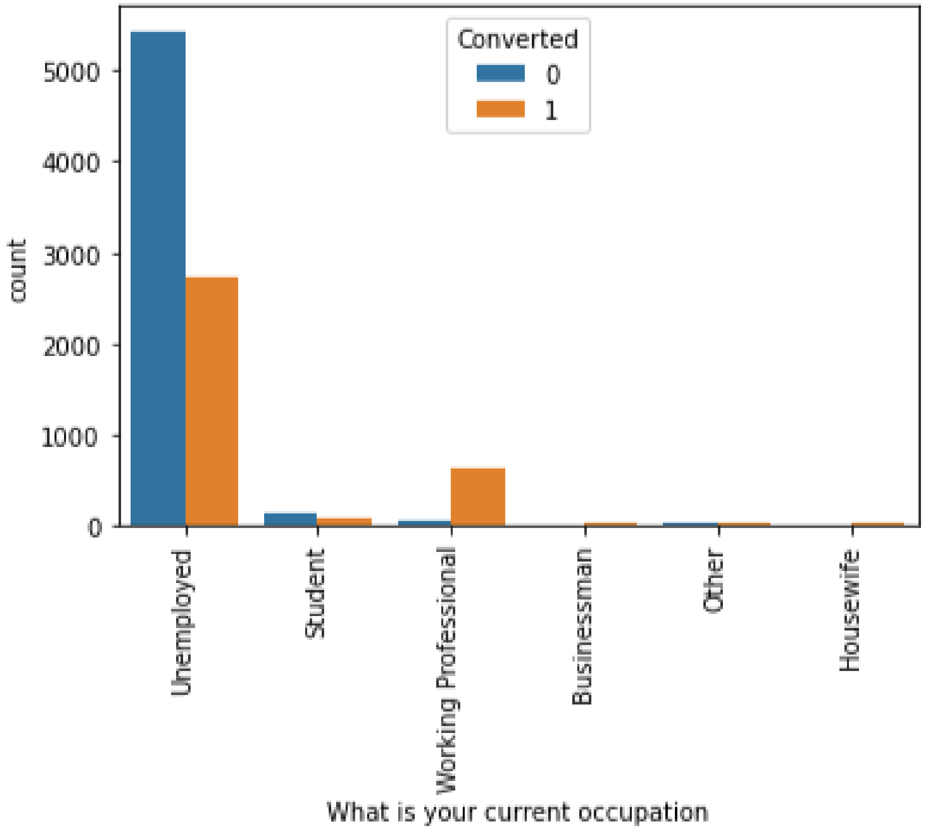


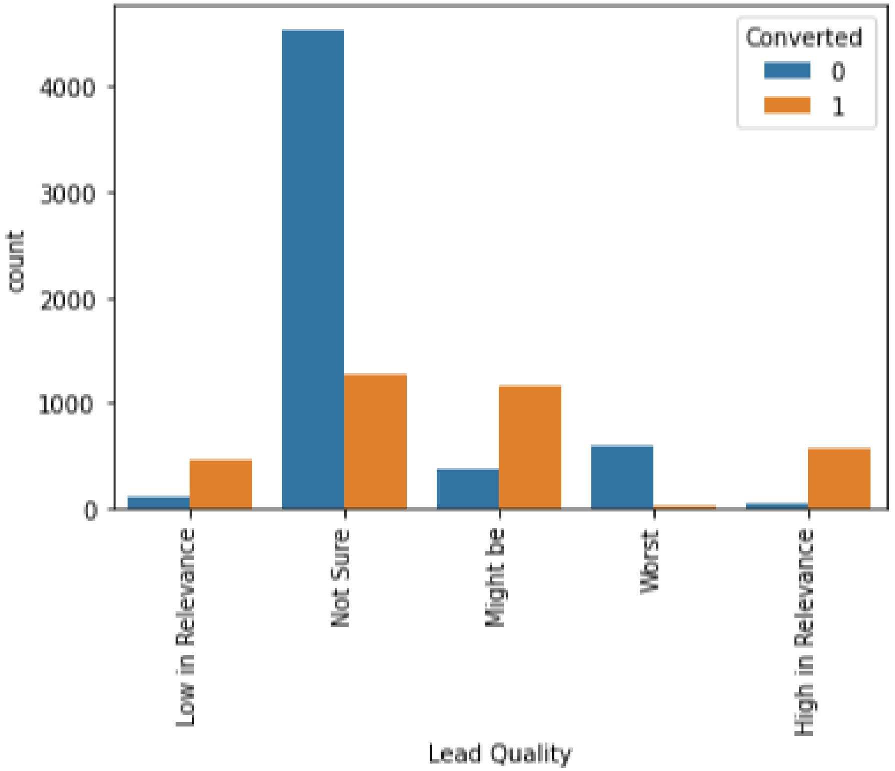
The conversion is highest with last activity as 'SMS sent' followed by 'Email opened' category.



The highest conversion specialization is to be found out. Finance, HR and Marketing are the next highest categories.

Data analysis :

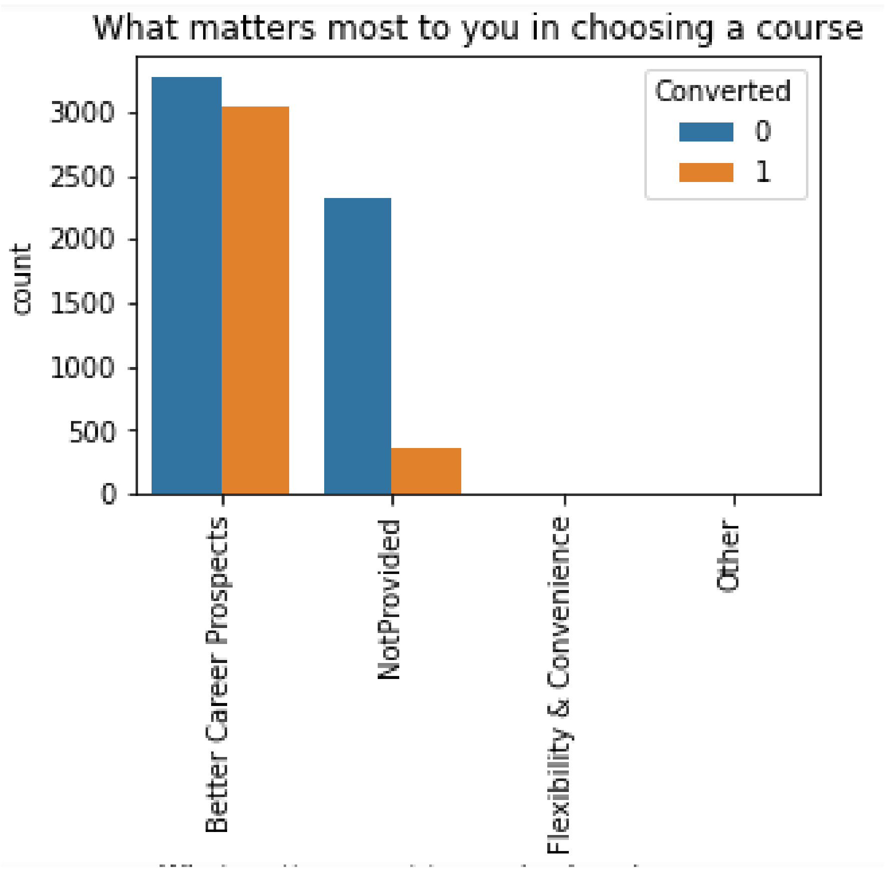


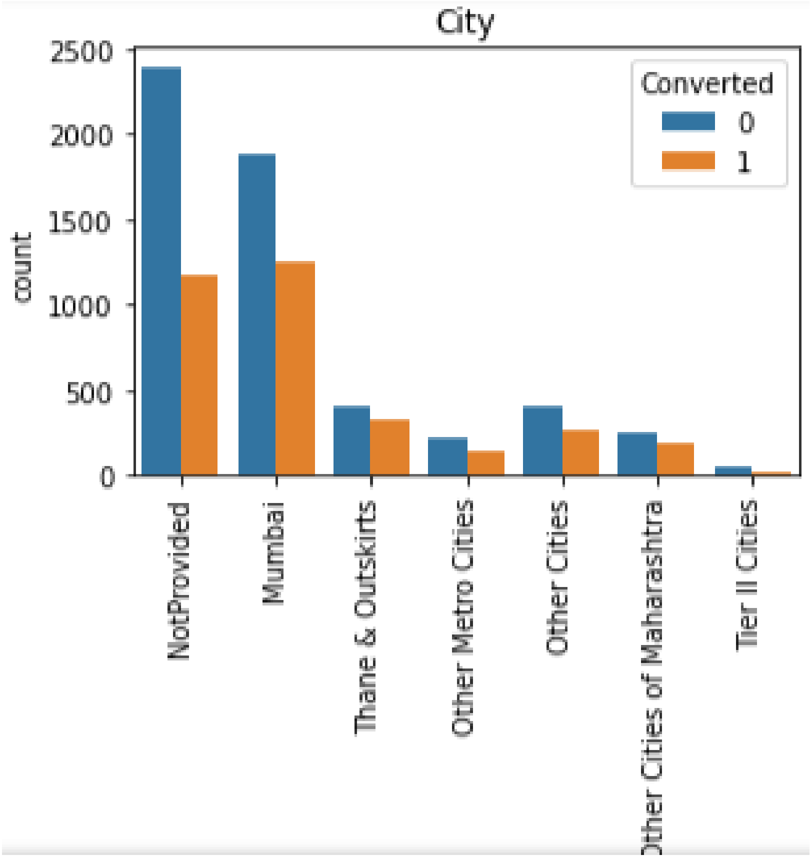


The conversion rate is highest in Working professionals category.

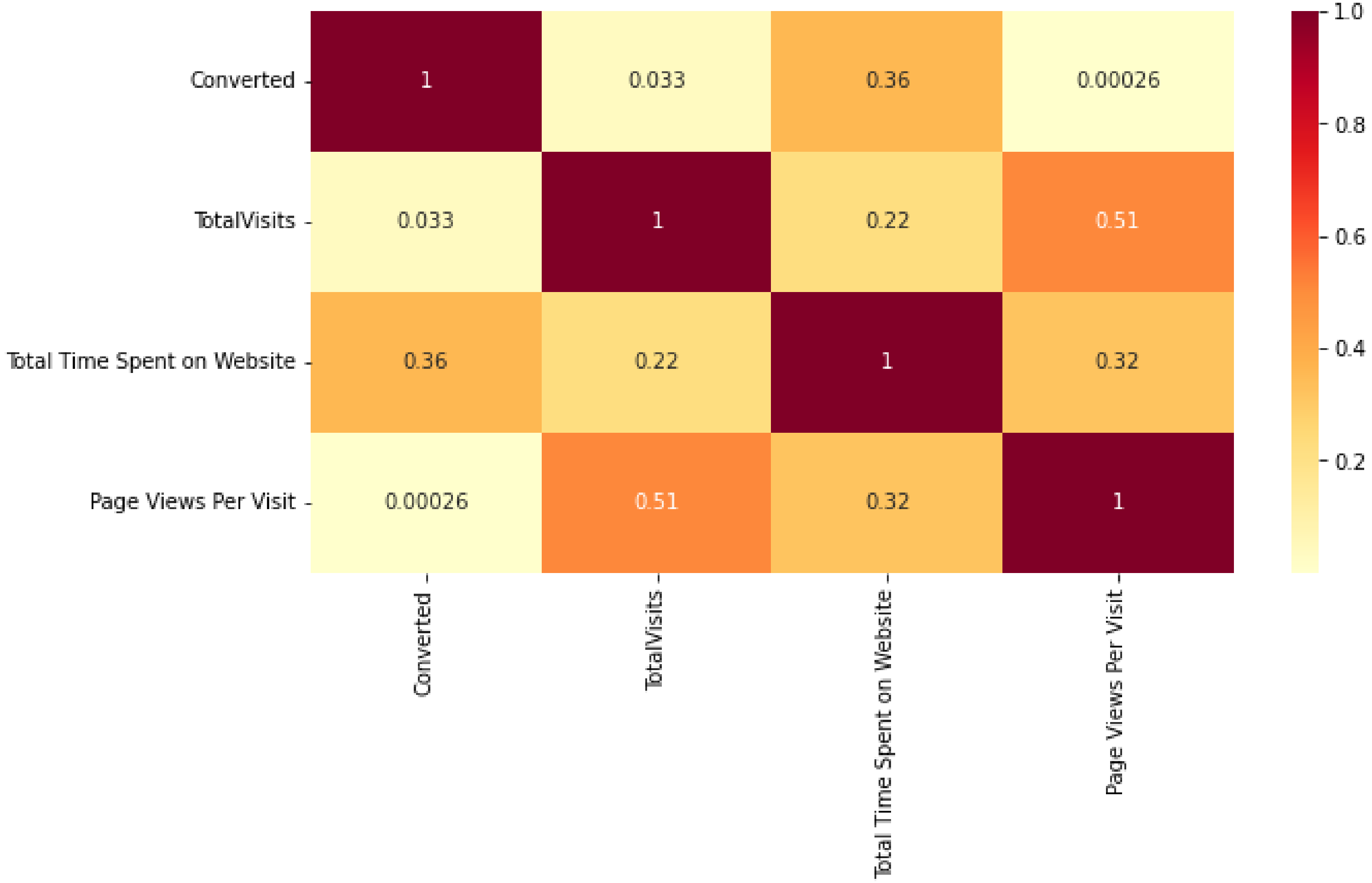
The conversion is higher in 'High / Low Relevance' and 'Might be' cases

Data analysis :



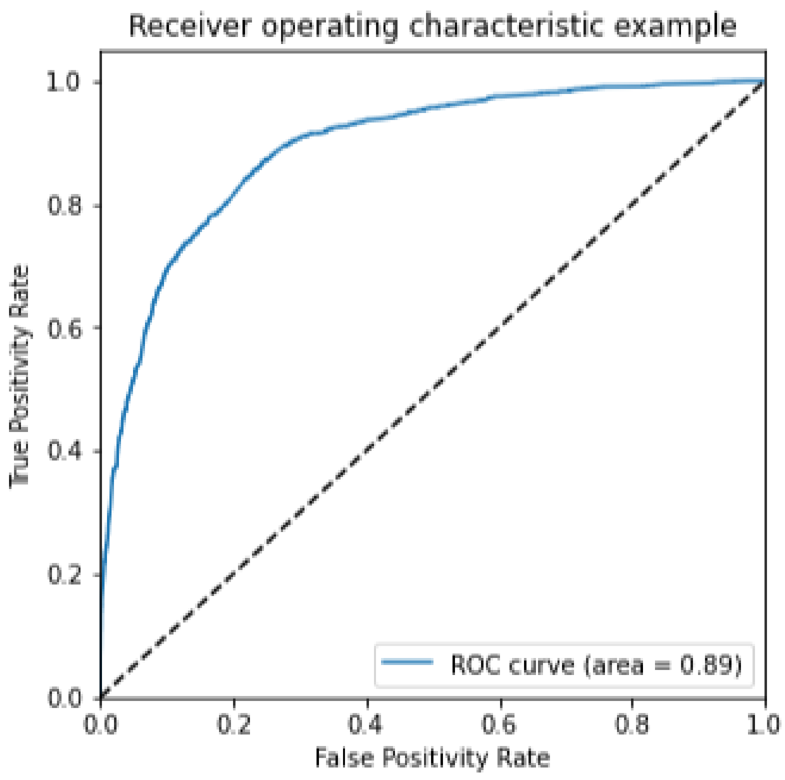


Data analysis :



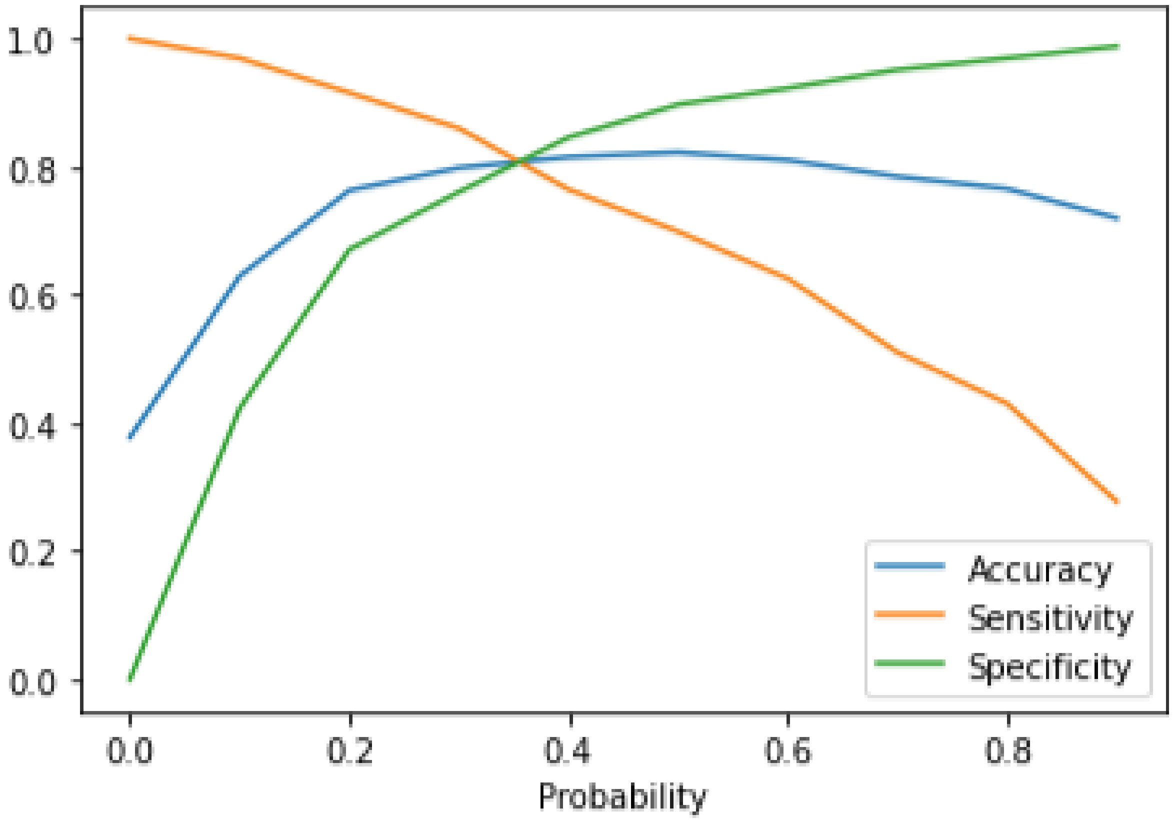
Model building :

- Split the data in 7:3 ratio for train & test.
- Used RFE for feature selection with 20 variables to selected.
- Removed variables with p-values higher than 0.05 and VIF higher than 5.



Area under the ROC Curve is 89%

Model prediction :



With the Cutoff set to 0.35 we have Accuracy, Sensitivity and Specificity around 80%

from the graph plotted we can see that the cutoff is 0.35

Conclusion

Following variables matter the most in deciding the potential lead in descending order.

- TotalVisits
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Referral Sites
- Lead Source_Welingak Website
- Last Activity_Email Bounced
- Last Activity_Olark Chat Conversation
- Last Activity_Unsubscribed

Conclusion

The below steps should help to reach out to the maximum potential leads with minimum sales executive to improve the conversion ratio.

- Make the homepage of the website attractive and information enough to grab people's attention.
- Target working professionals.
- The company can use of automated tools like emails / text / chatbots to carry out the communication part with minimal human intervention.

Thank you