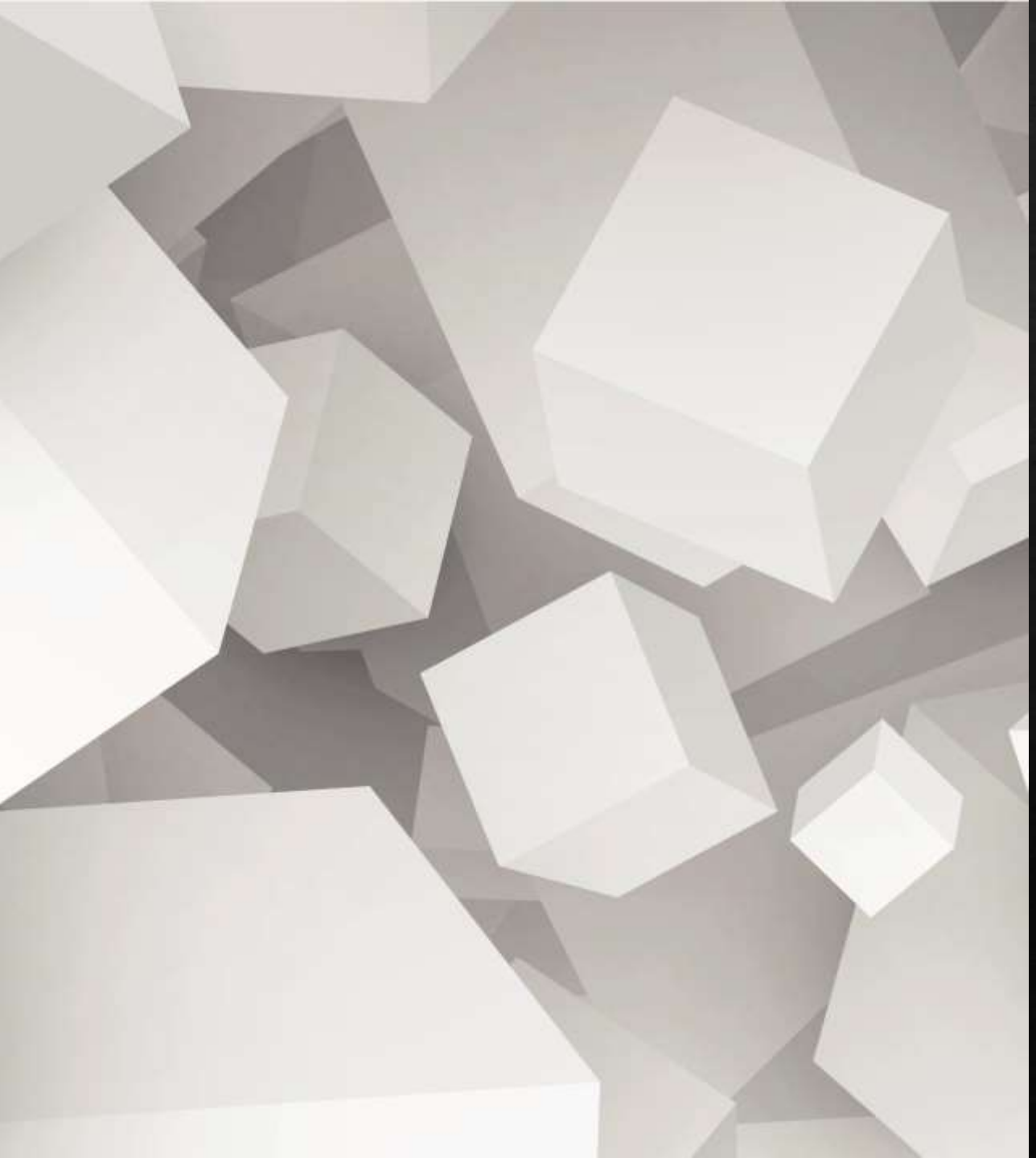
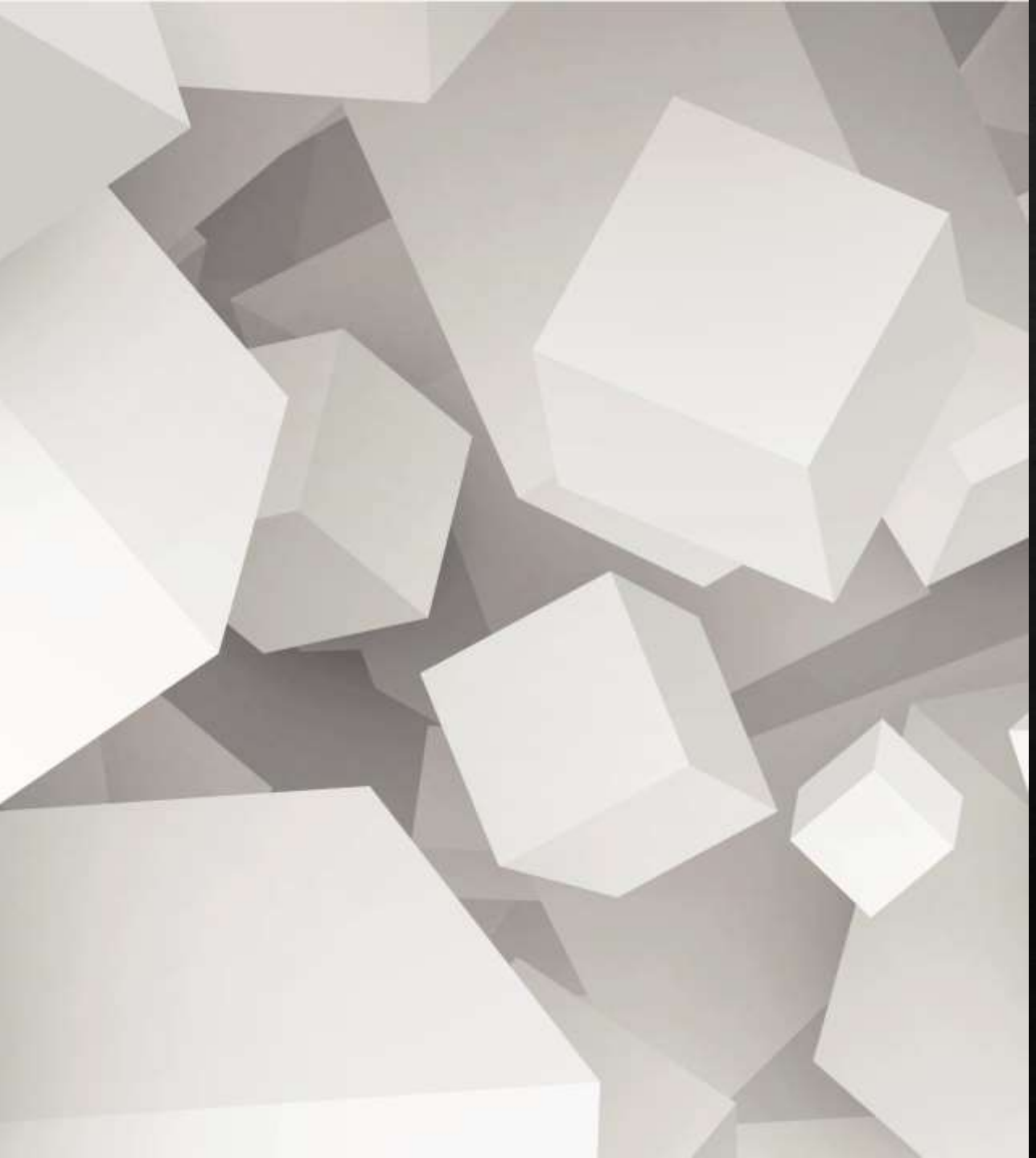
Lead Score Case Study



Problem Statement

- An online education company “X Education” sells online courses to industry professionals.
- X Education gets a lot of leads but the conversion rate has decreased significantly.
- To make the process smooth, the company wants to identify the parameters that would help them to find potential Leads.
- If they find these potential Leads, they could save up on a lot of resources and solely connect with them rather than connecting with everyone.



Business Objective

- The company wants to know potential leads also known as “Hot Leads”.
- For that, they have asked to build a model that would identify Hot Leads based on certain parameters,
- They also want the model to be future proof so that any future changes does not hamper the model.

Methodology

- Data preparation
  - Checking for duplicate data.
  - Checking for all the null and empty values, Replacing the empty spaces with nan values.
  - Dropping non relevant columns with large number of null values.
  - Checking and handling the outliers in the data.
- EDA
  - Univariate Analysis
  - Bivariate Analysis

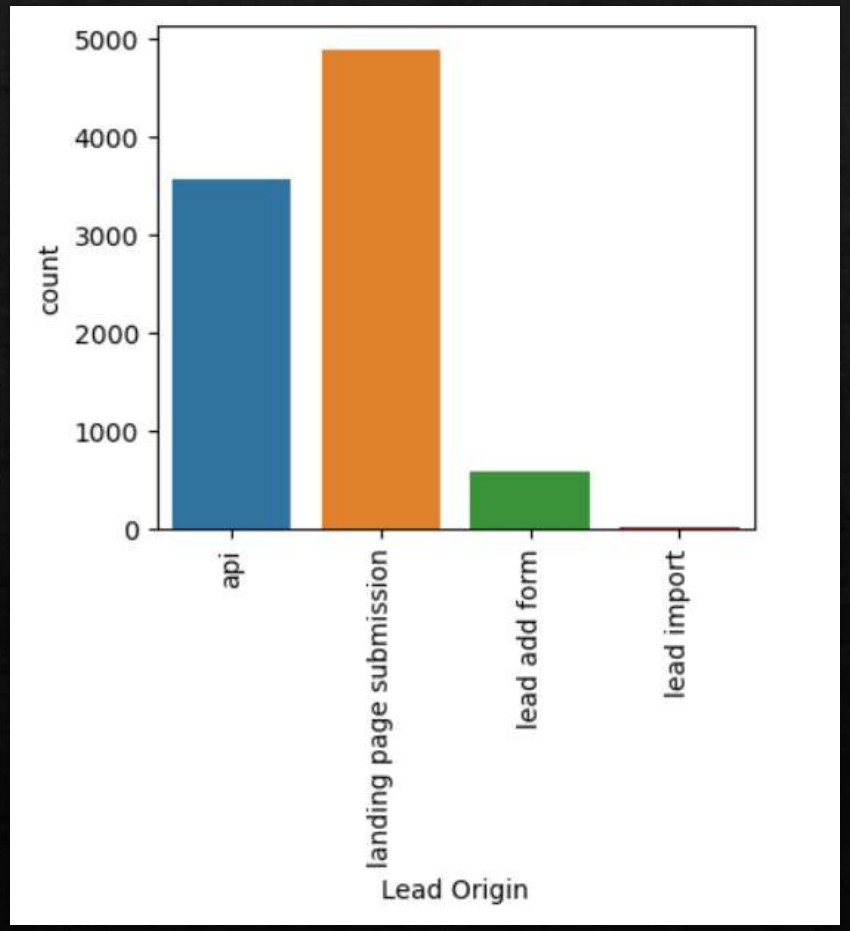
Methodology

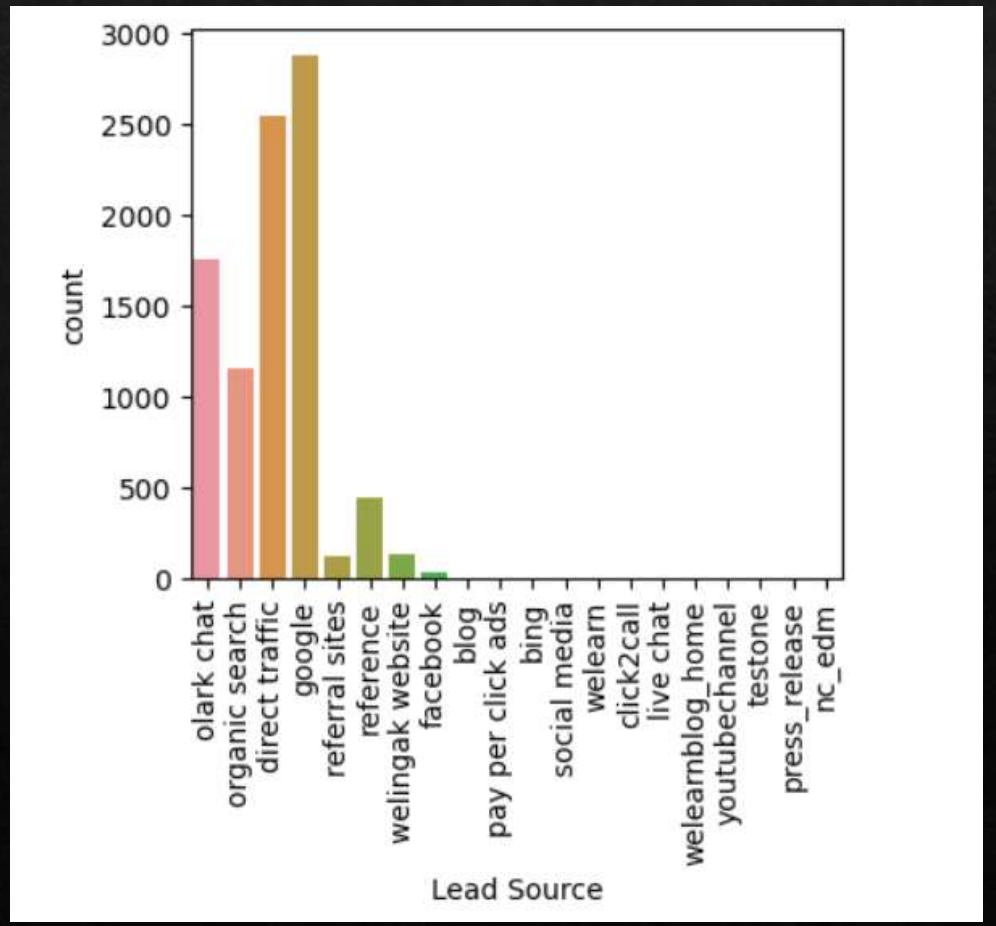
- Model preparation
  - Scaled the features and created dummy variables.
- Logistic Regression
  - Used Logistic regression as the classification technique for the model building and prediction.
- Evaluation and validation of the model.
- Conclusion.

Data preparation

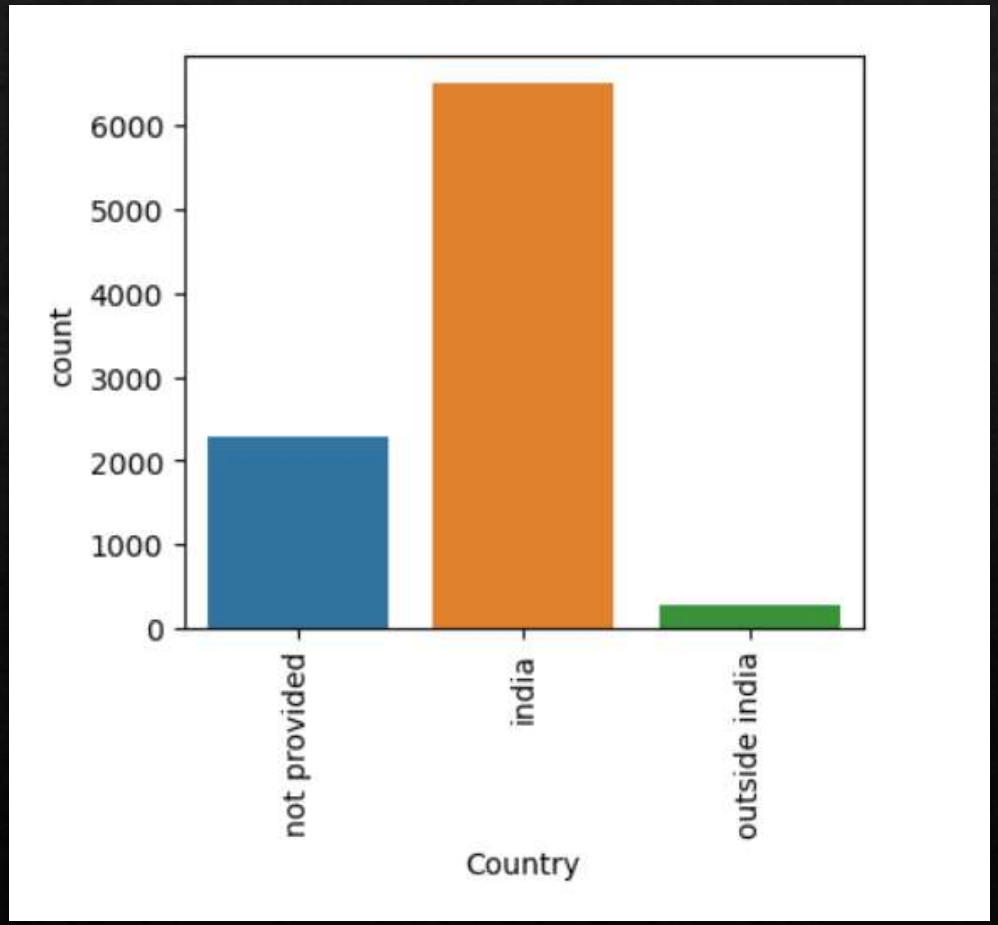
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’

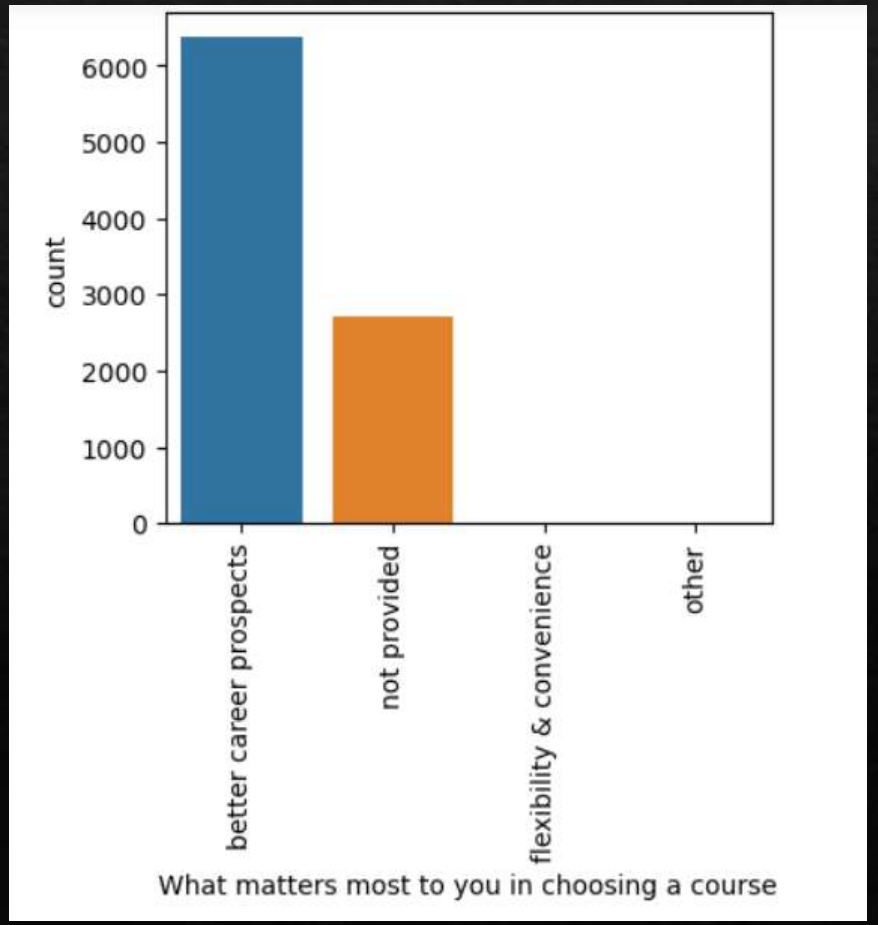
EDA – Univariate Analysis





EDA

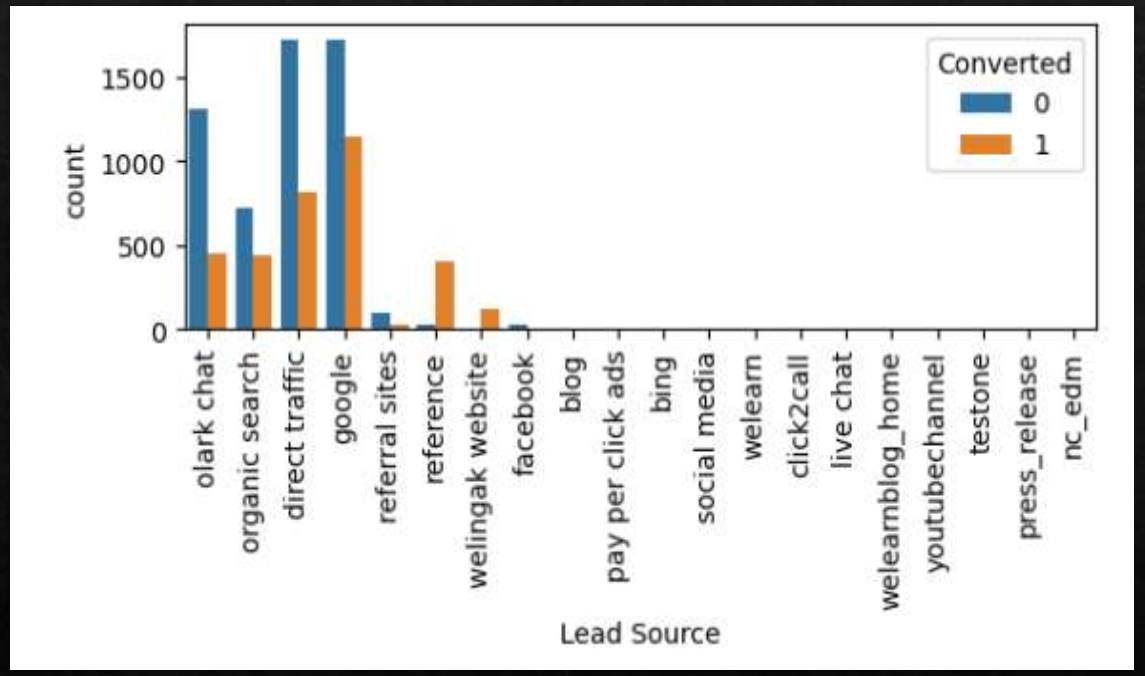

6000
5000
4000
3000
2000
1000
0
count
not provided
india
outside india
Country

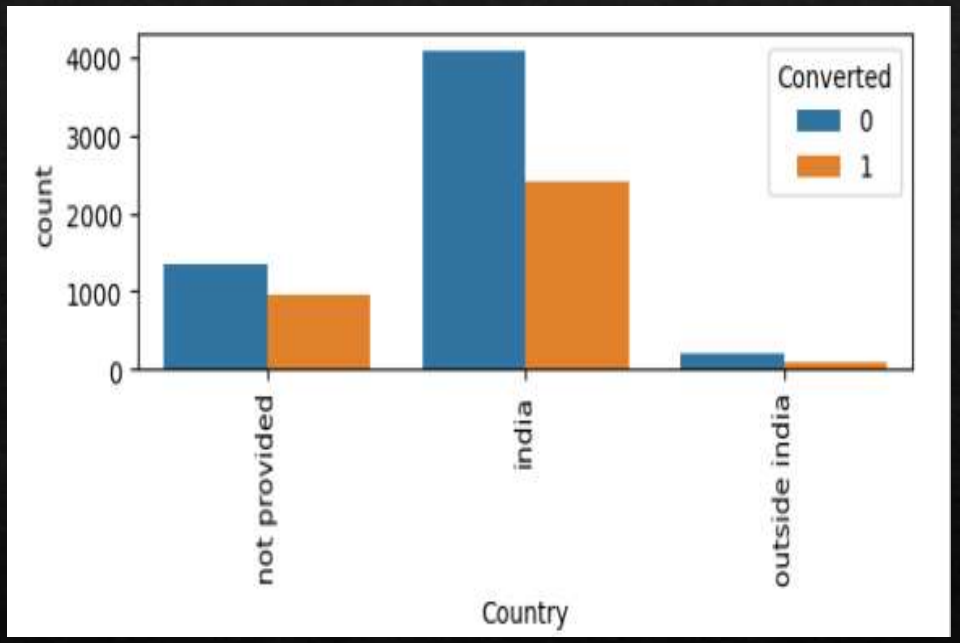

6000
5000
4000
3000
2000
1000
0
count
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course

EDA

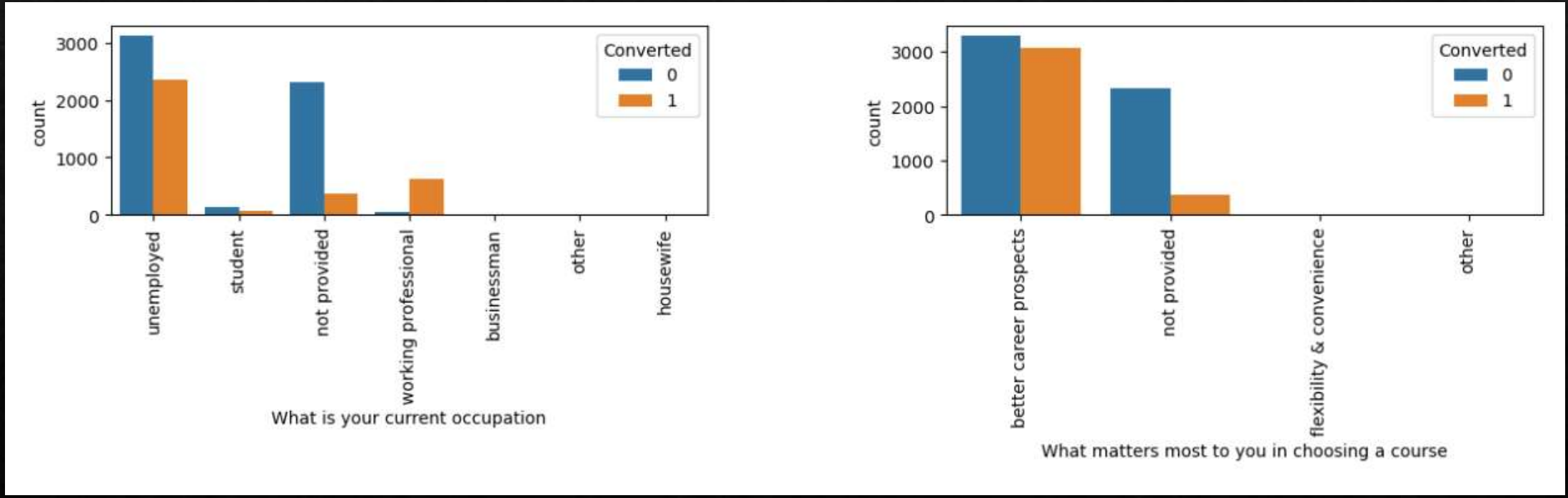
- Majority of the customers submitted their forms on the landing page
- Majority of the customers are sourced out from Google
- 70 to 80 percent customers belong to India
- Majority of the customers have completed their educationa and are choosing this course just for better job prospects

EDA – Bivariate Analysis





EDA – Bivariate Analysis



Model Preparation

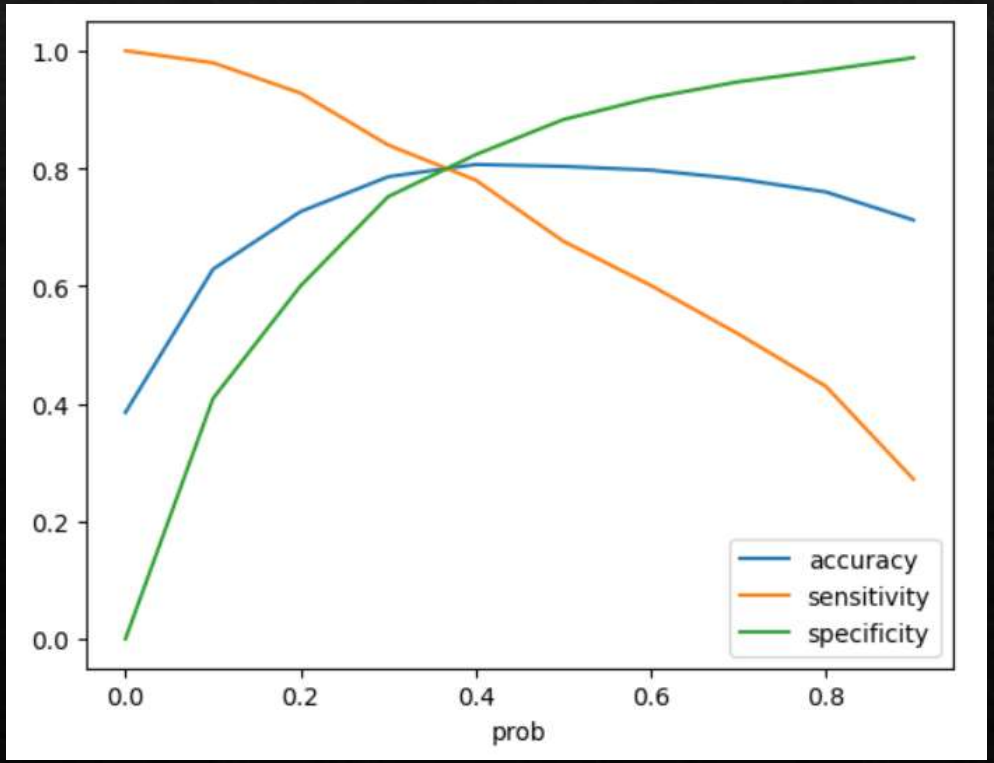
- The numeric variables in the dataset namely “Lead Number”, ‘Converted’, ‘Total Visits’, ‘Total time spent on website’ and ‘Page view per visit’, were normalized using MinMaxScaler.
- Dummy variables were created for the categorical variables.

Model Building

- The data was split into training and testing dataset (training: 70% and testing: 30%).
- We used RFE for feature selection as it was really hard to make out the correlation between the variables after creating the heatmap as there were so many variables.
- Ran RFE for 15 variables as the output.
- Built the model by checking the P value and the VIF and removed the variables with a high P value or high VIF.
- Lastly, made the predications twice by 2 different calculated cutoffs on the test dataset and the overall accuracy came out to be 81%.

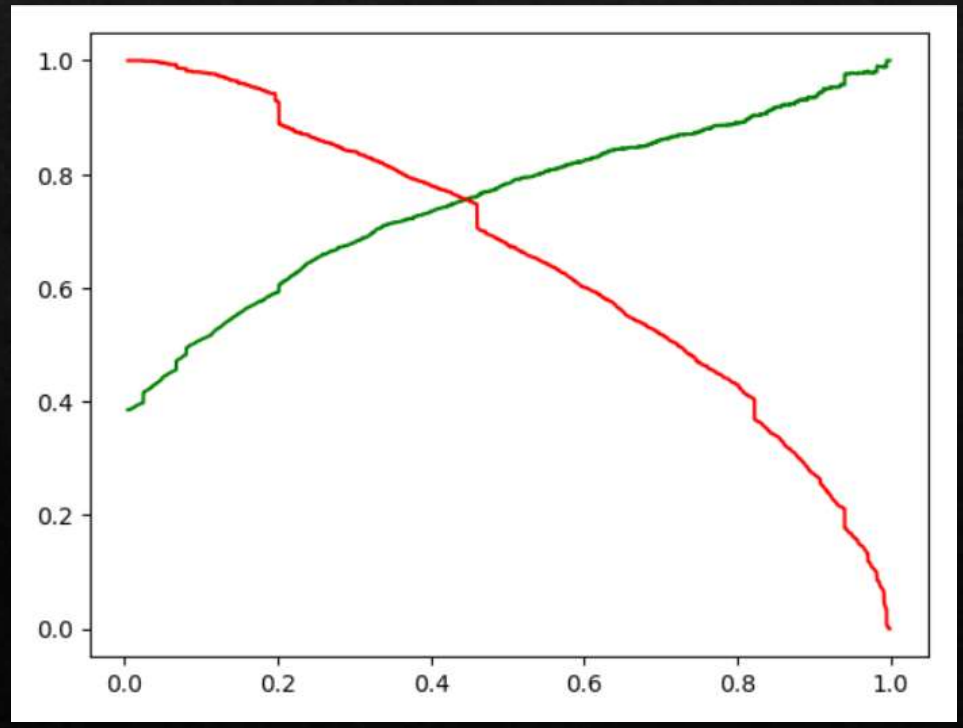
Model Building

- The cutoffs were decided by the ROC curve, where the cutoff came out to be 0.37.



Model Building

- Once the ROC cutoff was calculated, the precision and recall was also calculated and the new cutoff came out to be 0.43.



Conclusions

- The conclusions that came up after building and evaluating the model are:
  - The total number of time spent on the website is a primary parameter.
  - If the leads are interested they will add their info directly on the form.
  - The hot leads usually come through direct traffic and if they search organically.
  - Majority of the hot leads are working professionals.
- Keeping these in mind the company X Education can differentiate the hot leads form the normal leads and save up a few resources and expect a rise in their lead conversion.