Lead Score Case Study

BY

Ajay jain & SAPTARSHI

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Solution Methodology

- **Data cleaning and data manipulation.**

  1. Check and handle duplicate data.

  2. Check and handle NA values and missing values.

  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.

  4. Imputation of the values, if necessary.

  5. Check and handle outliers in data.

- **EDA**

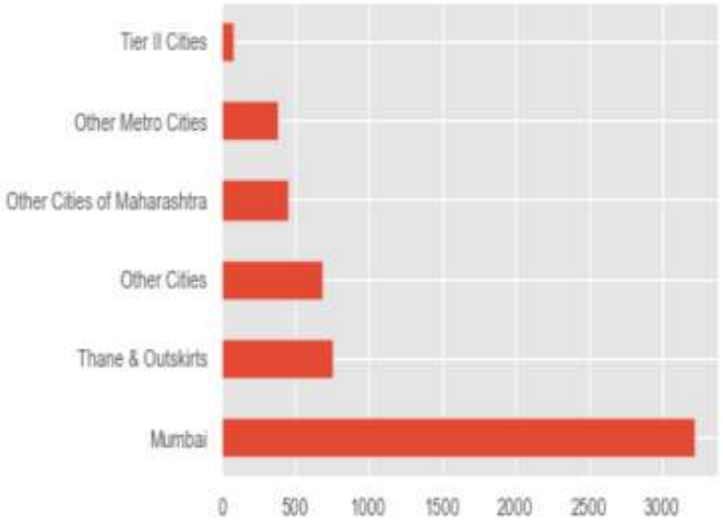
  1. Univariate data analysis: value count, distribution of variable etc.

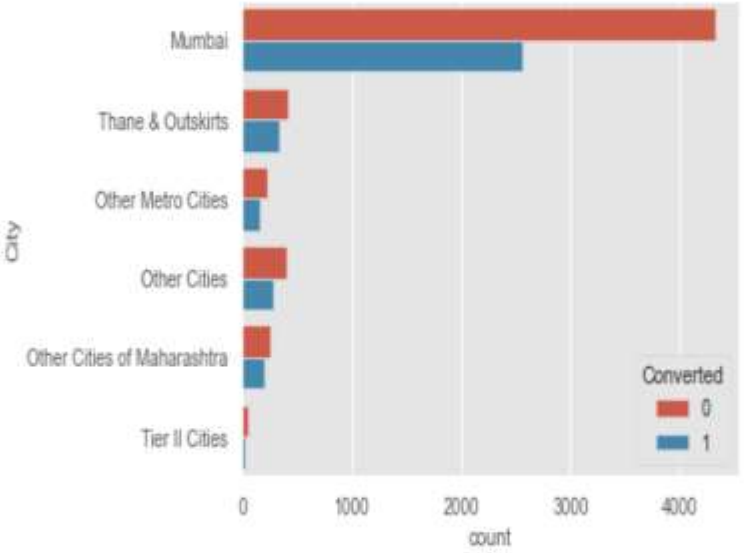
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

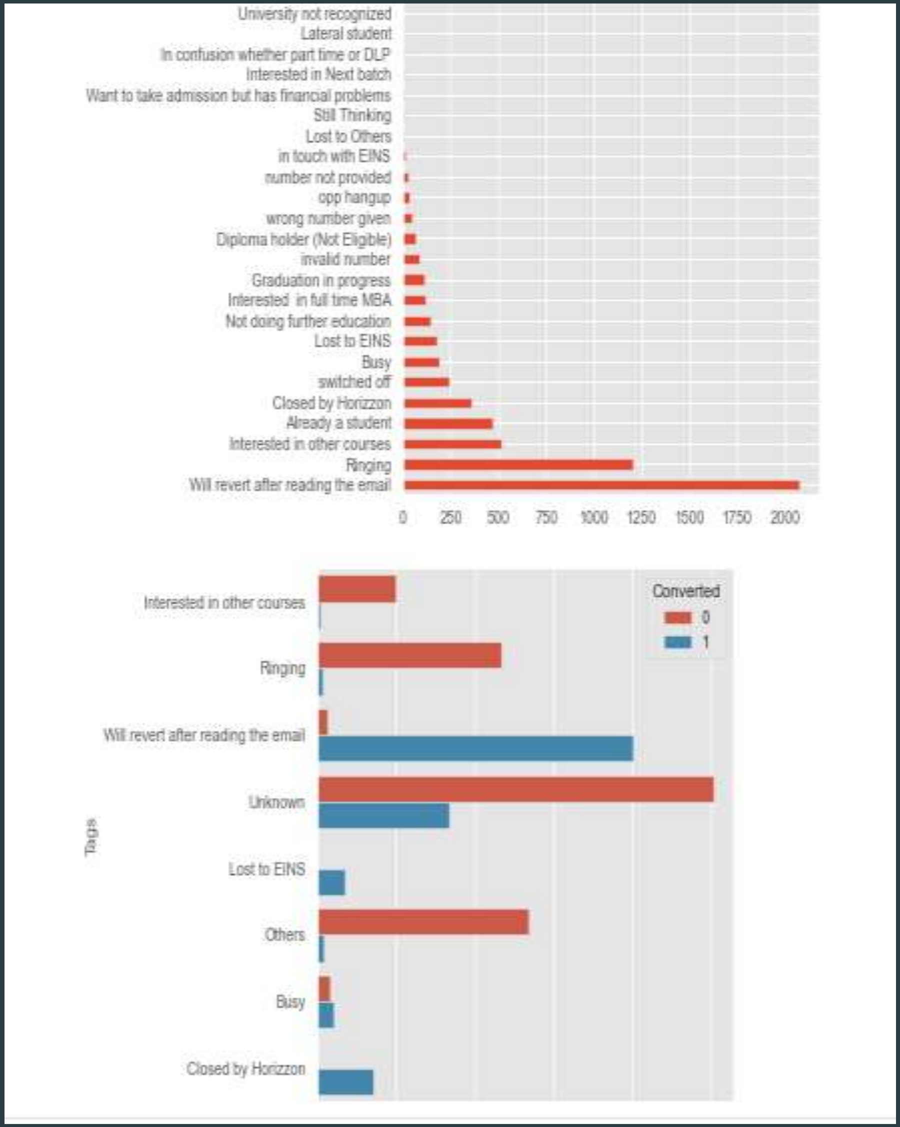
Distribution among different Cities

- As we can that Mumbai is at par the highest number of records .We'll replace nulls with Mode-'Mumbai'
- If Business involved, we can consult with them as well.
- For now we'll keep the feature, if model prediction won't depend on city, we will drop it then



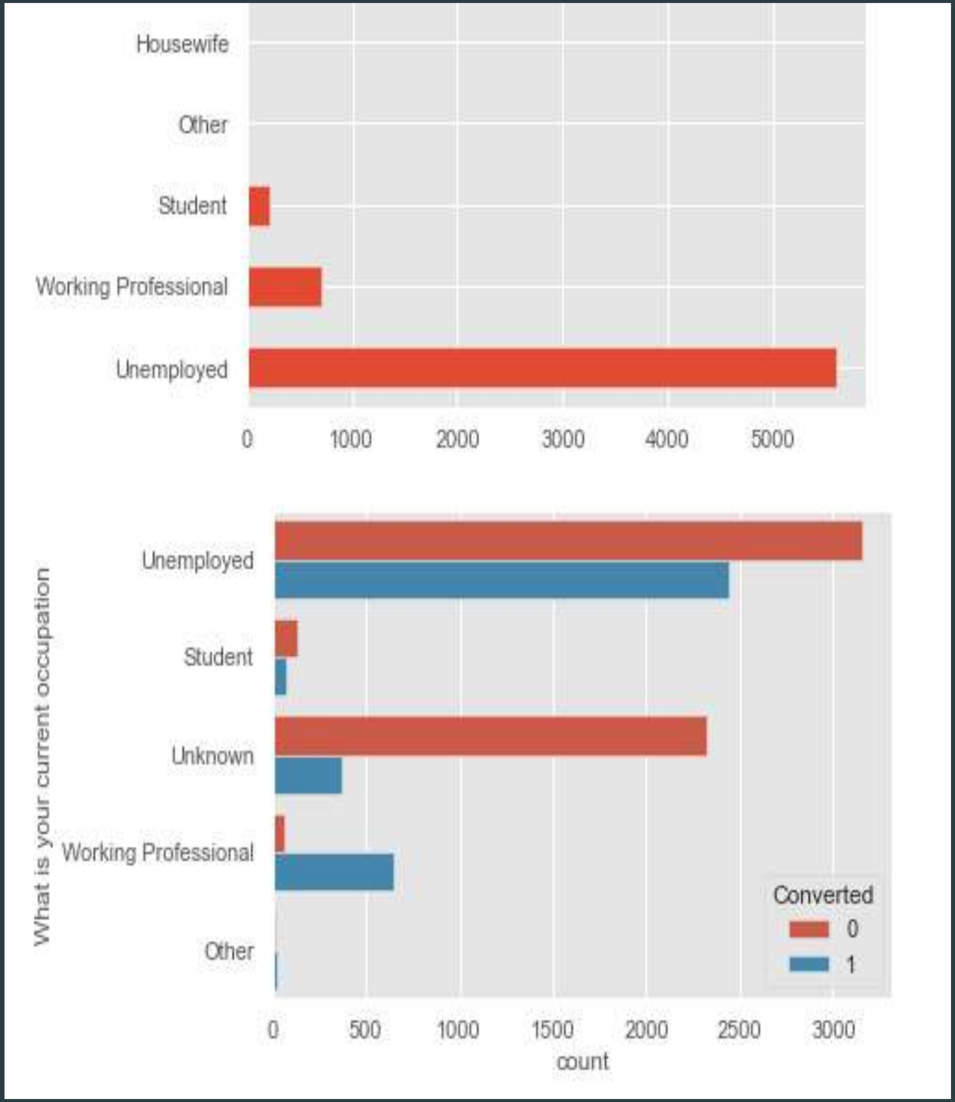


Distribution among different Tags



- 'Will revert after reading the mail' (2000 approx.) and 'Ringing'(1250) are not so far part
- Replacing 36% null entries to mode may cause issue and can make the dataset bias. Hence best to create another subcategory as Unknown
- Also, as there are many subcategories in Tags column, we'll club them together
- “Lost to EINS” , “Will revert reading mail” customer statuses are having highest conversion rate. They can be good predictors.

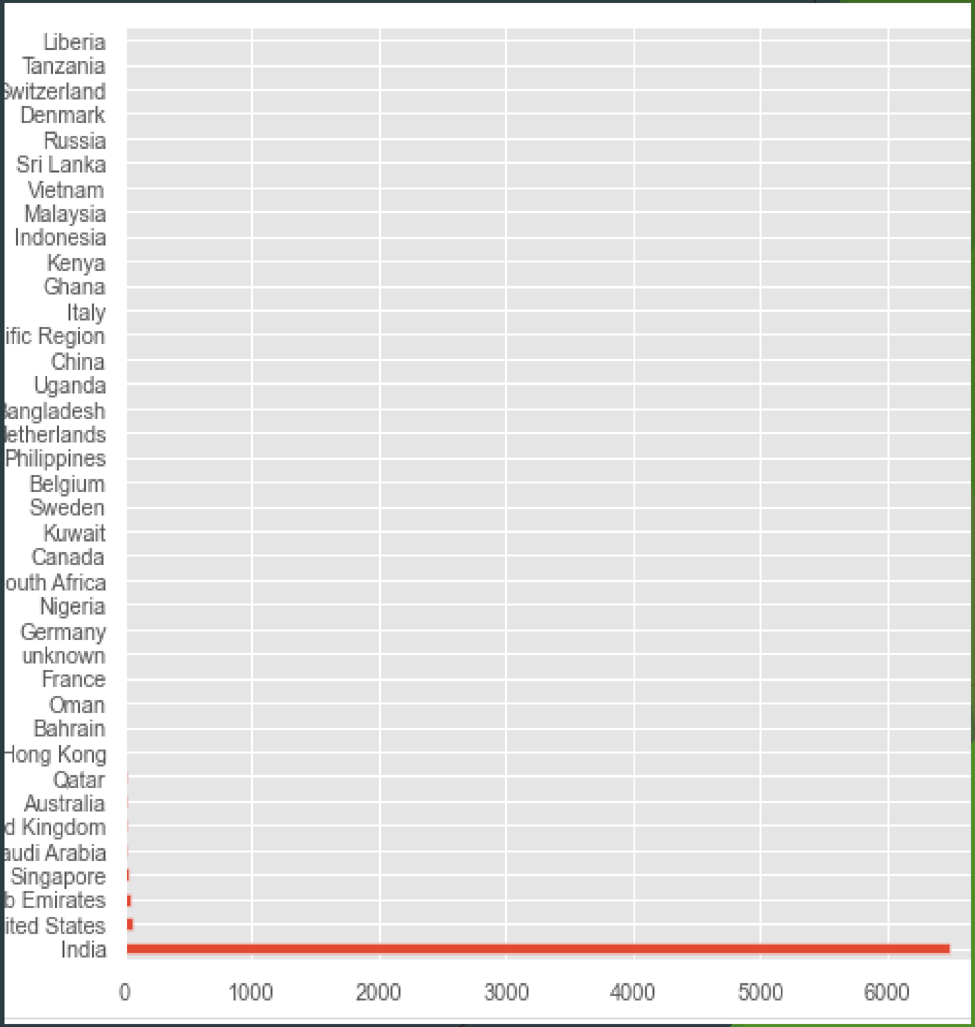
Distribution among Occupation



- Clearly 'Unemployed' is having 5k+ entries in the dataset hence making a large difference with other subtypes.
- We can replace 29% of the null values to mode ie, Unemployed
- Also, we can club Housewife and Businessman to Other as their count is very low.
- “Working Professional “ and “unemployed “ subcategories have highest conversion rate, can be good predictors for modelling.

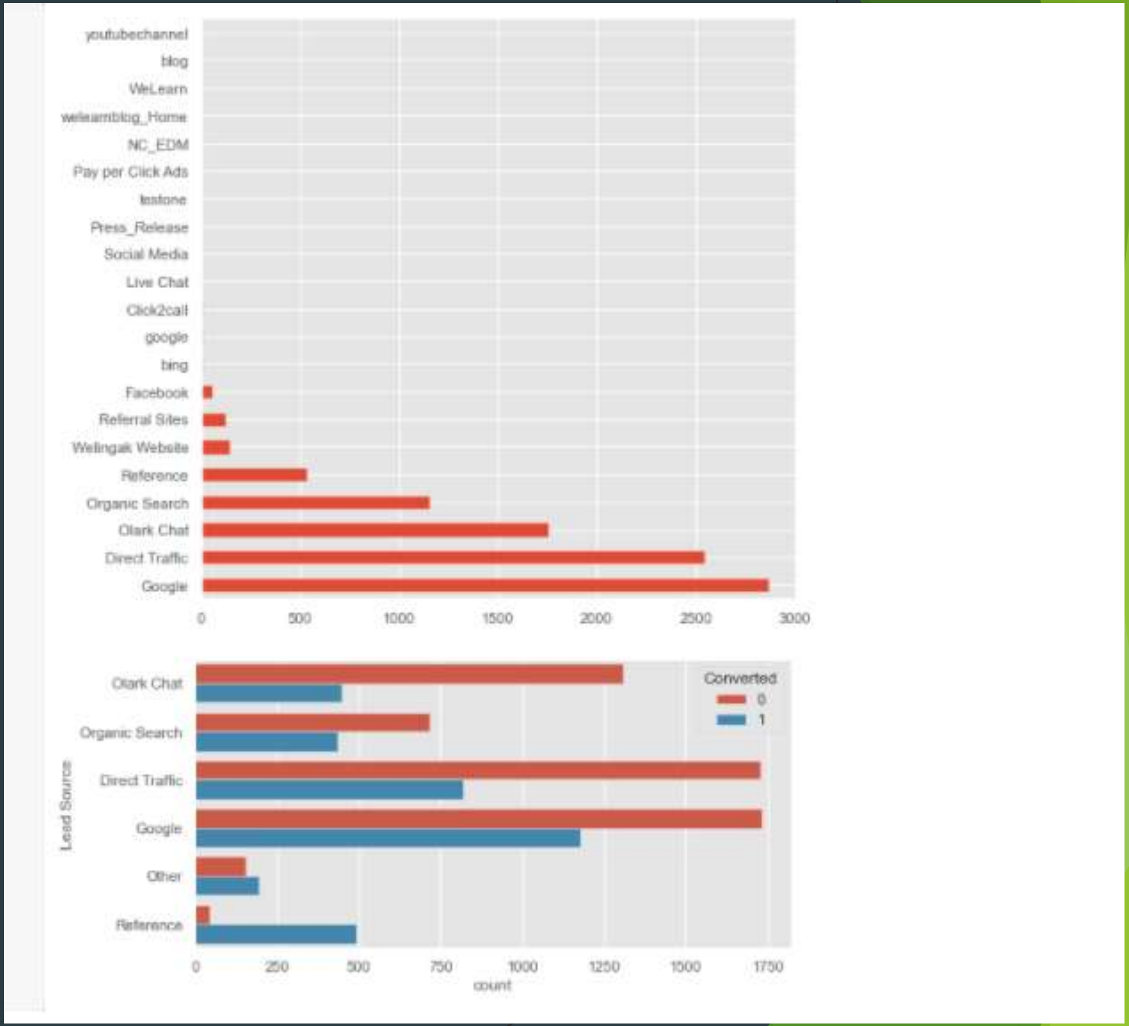
Distribution among Countries

- India clearly has the highest count of customers, hence Indian people would be the best to target. That is a fact and can be used for Business operations.
- As their is not much variance in the columns we would simply drop it.

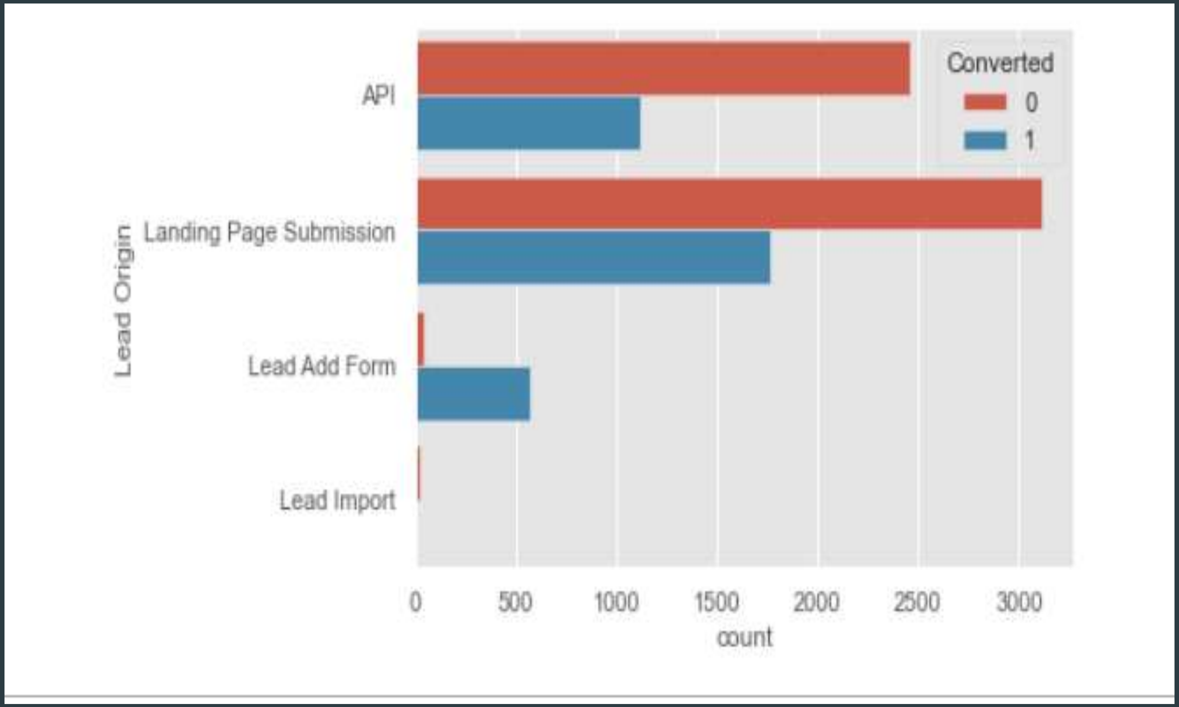


Distribution among different Lead Sources

- We will impute 36 null values with mode- Google.
- Also we'll bin the lowest count subcategories into Other
- “Google” and “Direct Traffic” lead source have highest conversion rates. Also the customers that have source as “reference” should be a target, as their conversion rate is high.

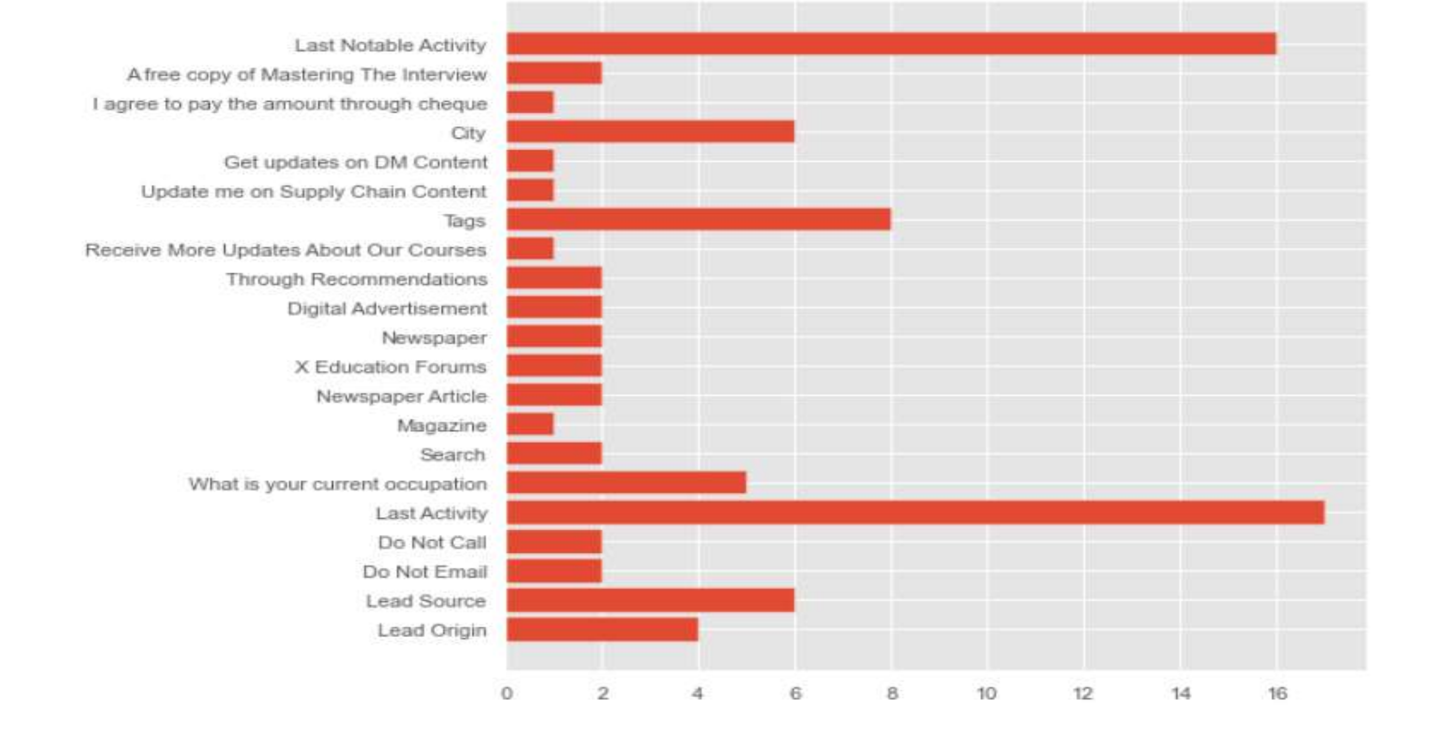


Distribution among different Lead Origin



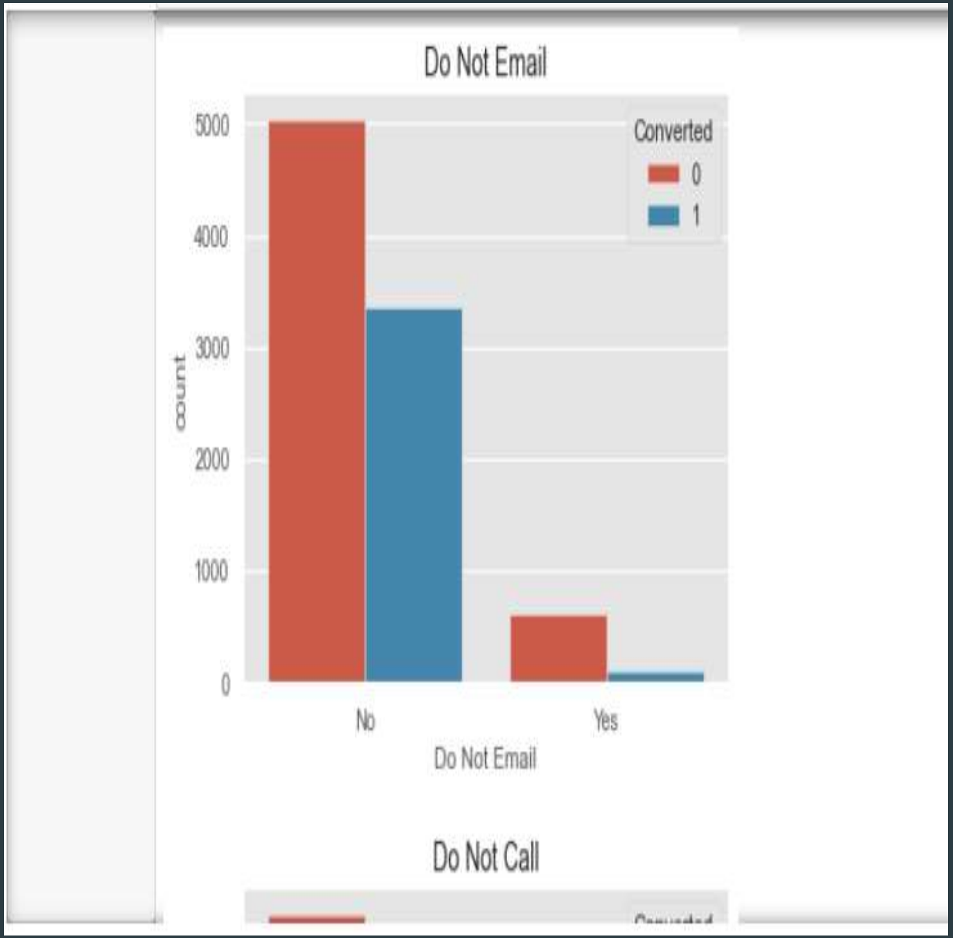
- Lead Add Form has the highest conversion rate while Lead Import has least. Customer filling out Lead Add forms can be a very good predictor and these customers should definitely be targeted .

Last Activity/Last Notable Activity



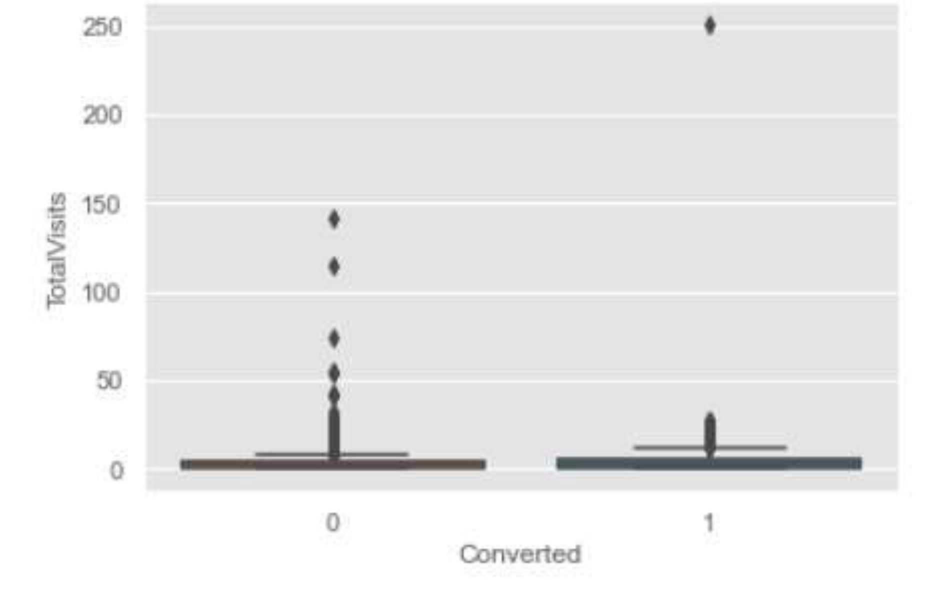
“SMS Sent” and “Email Opened “ have highest conversion rates. Also “Phone Conversation”/ “Unreachable “subcategories themselves have more convertors than non converters. Therefore if the customer’s last activity/ last notable activity is from any of the above , they are more likely to convert , and worth company’s effort.

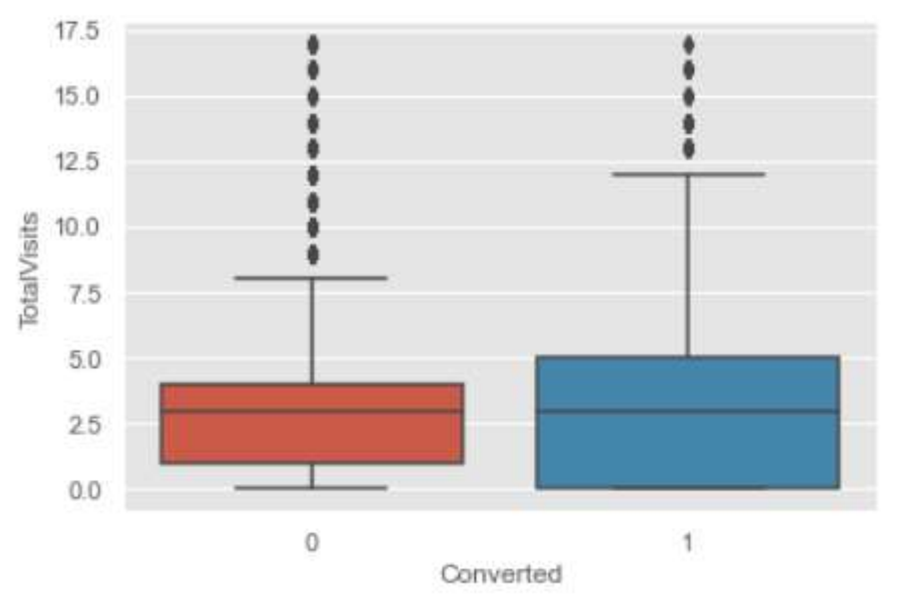
Do not call, Do not email



- As we can see columns-: 'Do Not Call','Search', 'Newspaper Article', 'X Education Forums','Newspaper', 'Digital Advertisement','Through Recommendations' do no have data for Yes sucategory- hence no varicance present in the columns. We will drop all these columns.

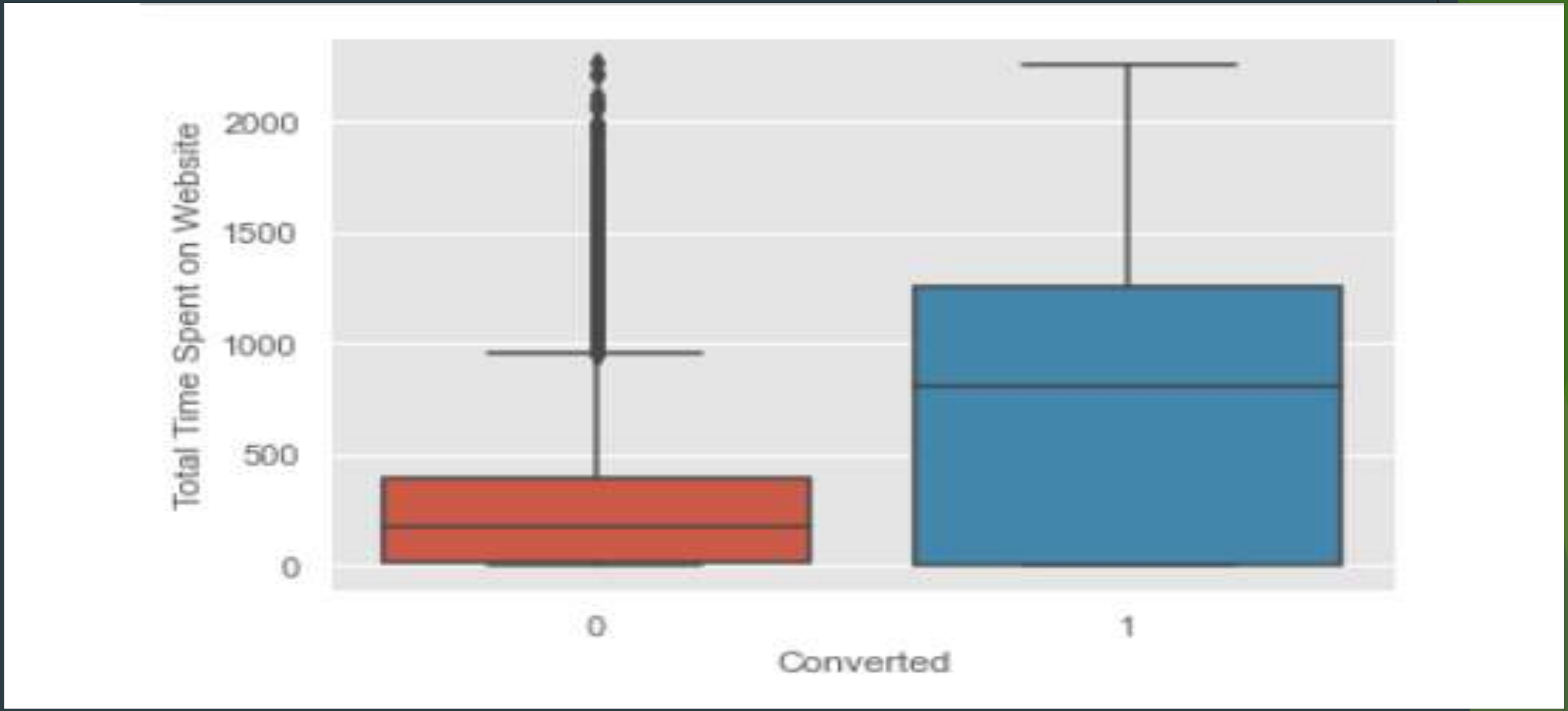
Total Visits





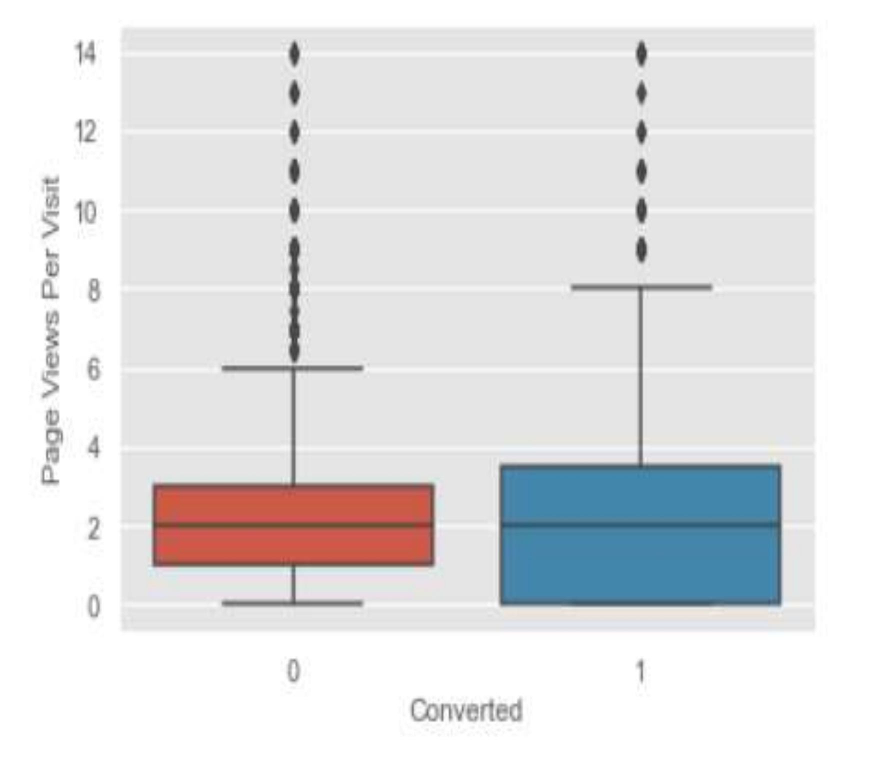
Higher distribution is present in converted class where Total visits varies from 0-5. This can account for the fact that customer is thinking of taking up the course and hence making sure if the investment would be right. These kind of people if talked through can be converted.

Total Time Spent on Websites



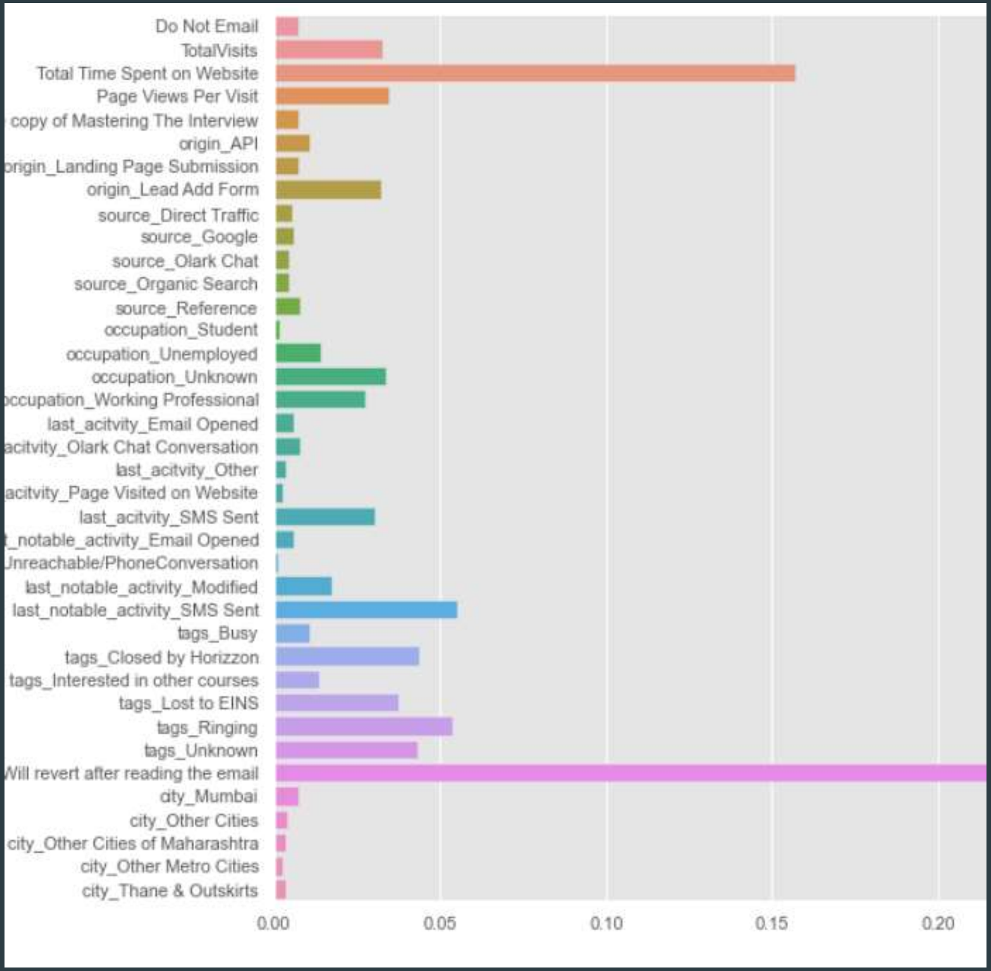
This is related to TotalVisits analysis above- if totalVisits increase then time spent on the website will also increase (generally). Hence we have higher distribution of converted people where time range is 0-1200.

Page Views Per Visit



- Median for both classes is almost same only, but distribution is higher for converted people in the same page views range. Many customer are present of converted class which have high page visits, this means that they explore the website to find what all perks/demerits are there for joining the course. Application usability and content would play a important factor to engage more people.

Feature Importance

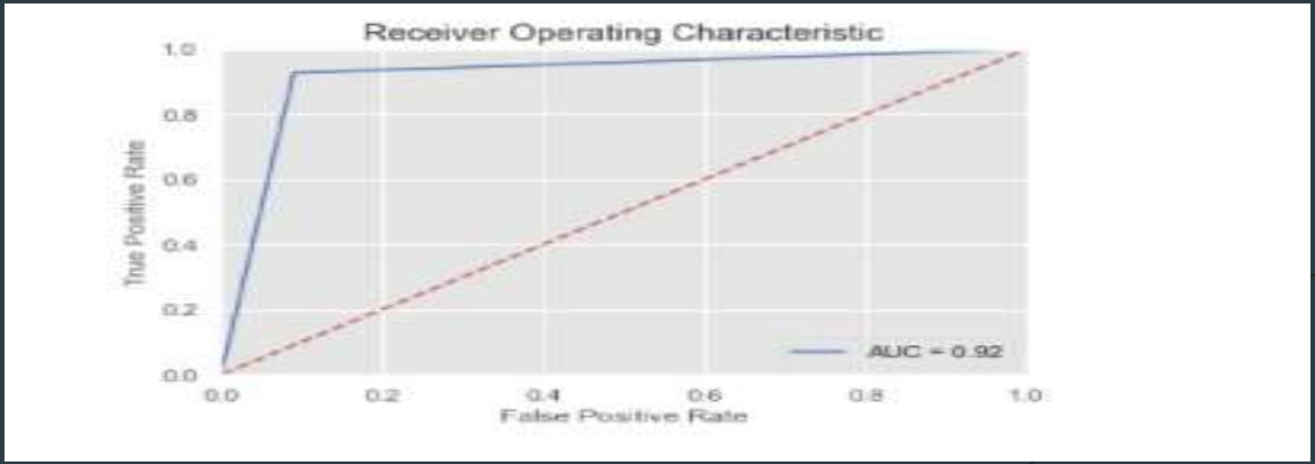


- TotalVisits
- Total Time Spent on Website
- Page Views Per Visit
- origin_Lead Add Form
- occupation_Unknown
- last_acitvity_SMS Sent
- last_notable_activity_SMS Sent
- tags_Closed by Horizzon
- tags_Lost to EINS
- tags_Ringing
- tags_Unknown
- tags_Will revert after reading the email
- occupation__Working Professional

Model Evaluation Metrics

- Accuracy: 0.91
- Sensitivity: 0.93
- Specificity: 0.91
- Auc score: 0.92
- F1 score: 0.89

Model seems to predict atleast 90% of the data correctly.



Model Predictor Features

- tags_Will revert after reading the email
- tags_Lost to EINS
- tags_Ringing
- tags_Closed by Horizzon
- last_notable_activity_SMS Sent
- origin_Lead Add Form
- Total Time Spent on Website
- Page Views Per Visit
- Tags_unknown
- Occupation_unknown

Apart from these unknown occupation and unknown tags features increase the model performance by 2% , but dropped them off from the final model because no business utility could be generated from them.

Observations and Recommendations

- Out of all the model predictor variables -

  tags_Will revert after reading the email, tags_Lost to EINS, and tags_Closed by Horizzon have highest coeffiecients and hence impact the predictiive power a lot.

- Company should focus on below features the most:

  - The customers who have a current status (Tags) as "Will revert after reading the email", "Lost to EINS","Closed by Horizzon" are most likely to convert and buying the course.
  - Total Time spent on Website is another feature which has a lot of impact on converting a lead. So if a customer is putting effort in exploring the whole website, they can be enticed by either sending mails/phone calls, this would result in higher conversion rates
  - One thing to note is tags_Unknown and occupation_Unknown impact the model significantly, hence just leaving out the customers who have not filled in any of the above categories doesn't mean that they are a less likely candidate. For these type of customers we can closely monitor other features like Total time spent on website/ page views/last_notable_activity and accordingly should take action.

Observations and Recommendations

- Customers who have lead origin as Lead Add Form are also more likely to be converted.
- If bandwidth available, company should focus on all the model features, this will ensure accuracy of approx 90% . They can also focus on customers that had lead source as referenced as they also have high conversion rates.
- When the situations are too tight it is best to focus on the customers who have a current status (Tags) as "Will revert after reading the email", "Lost to EINS", "Closed by Horizzon" as they are most likely to convert and buying the course. Team can also monitor the customers total time spent on website and filter out the ones having time range within 1500.