LEAD SCORING ASSIGNMENT

By : Cheshta,Joel,Sandeep

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

STEPS FOR PROCESSING

- Data Cleaning
- Data Preparations
- Model Building
- Model evaluation

DATA CLEANING

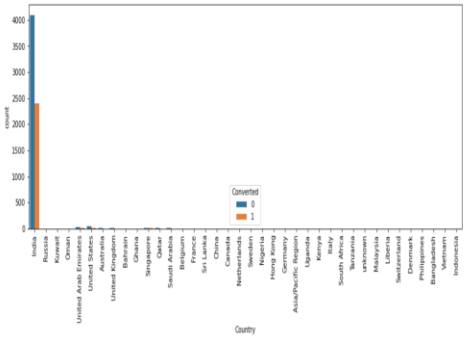
Checking for ‘target’ variables and duplicate values of id

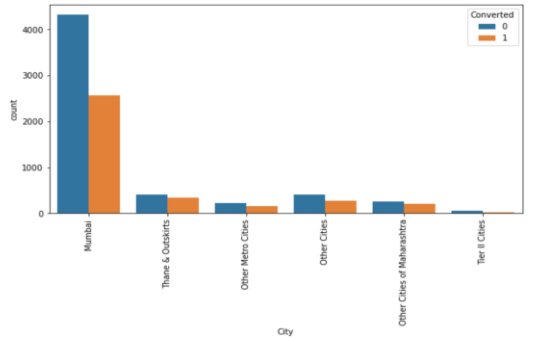
- Handling the Select levels and converting them to Nan
- Dropping high percentage of missing values
- Checking unique values and plotting them
- Prospect ID & Lead Number are two variables that are just indicative of the ID number of the Contacted People & can be dropped.

EDA- UNIVARIANT ANALYSIS

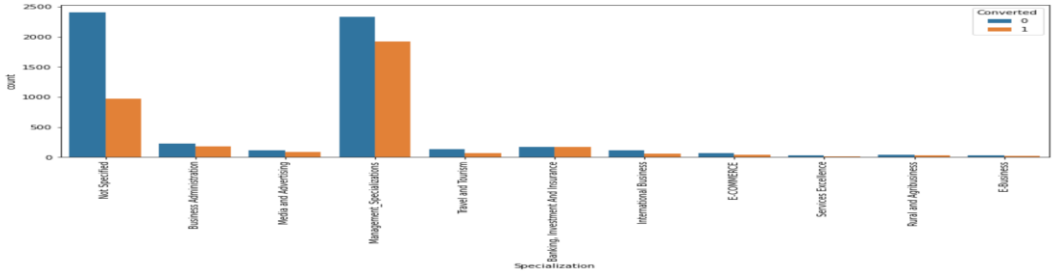
What matters most to you in choosing a course, ‘nan’ are replaced with ‘India’ and rest are dropped

So is for City , nan is replaced with Mumbai

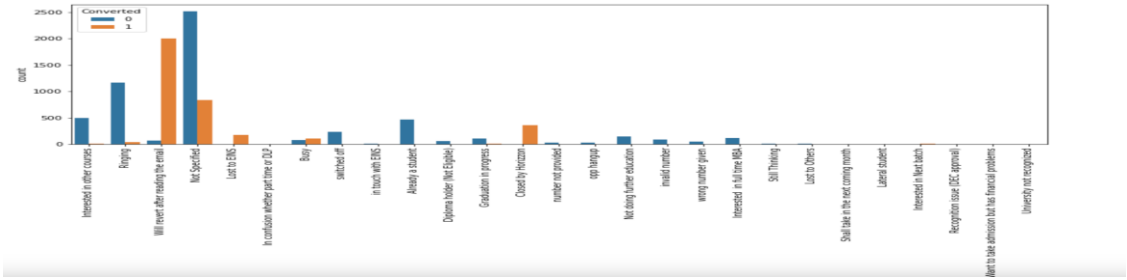




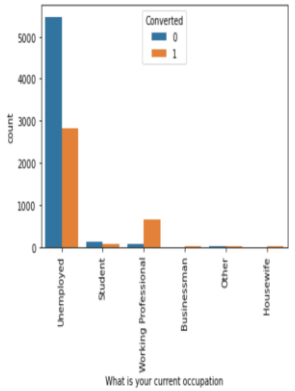
Splcialisation ‘nan’ are replaced with ‘Not Specified’ and we see that specialization with **Management** have higher number of leads as well as leads converted.
So this is definitely a significant variable and was not be dropped.



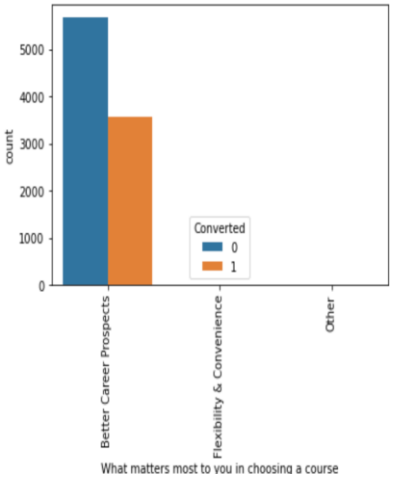
Tags , the nan are replaced with ‘Not Specified and then replacing tags with low frequency with "Other Tags”



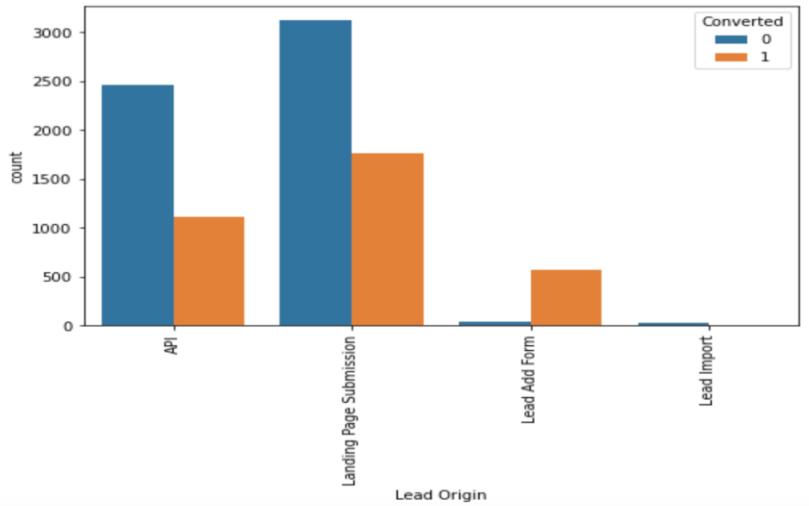
•For What is your current occupation, nan is replaced with Unemployed and we noticed :
Working Professionals going for the course have high chances of joining it.
•Unemployed leads are the most in terms of Absolute numbers.



•What matters most to you in choosing a course are choosing for Better Career Prospects , hence we can drop the coloumns



For Lead Origin , replacing values of nan and inferences :
•API and Landing Page Submission bring higher number of leads as well as conversion.
•Lead Add Form has a very high conversion rate but count of leads are not very high.
•Lead Import and Quick Add Form get very few leads.
•In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

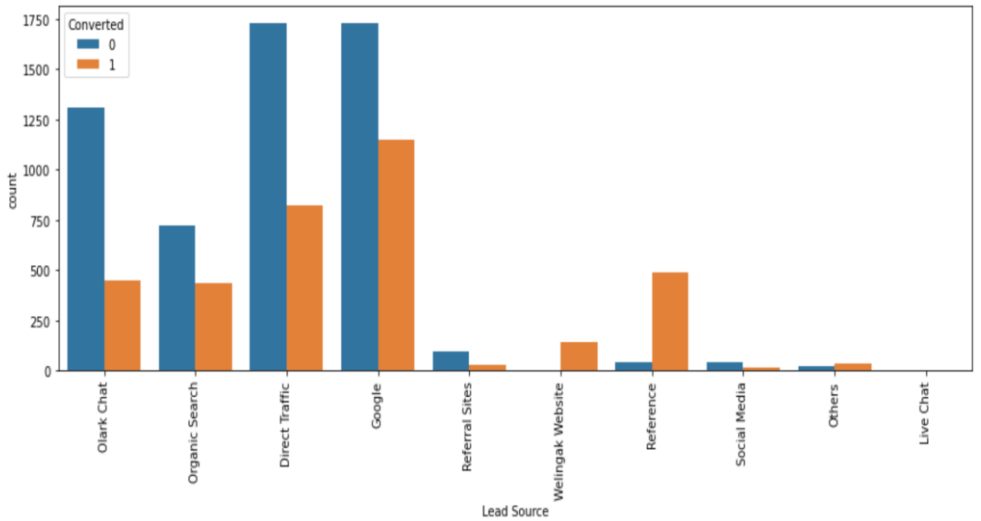


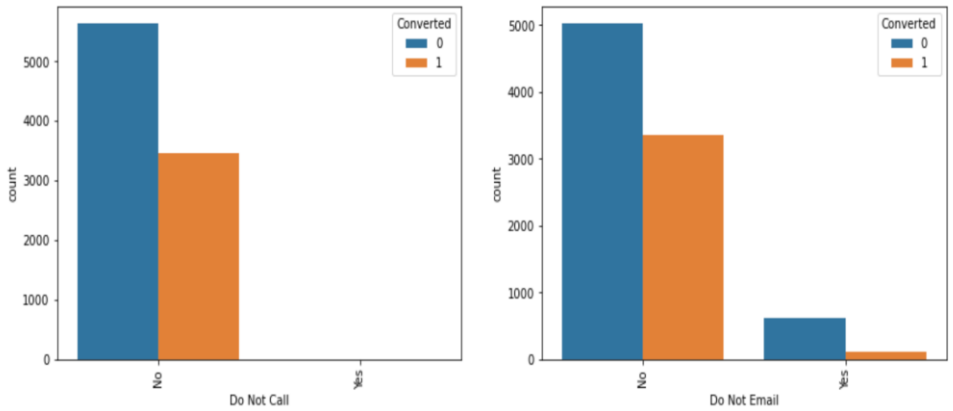
For Lead Source ,replacing ‘Nan’ Values and combining low frequency values with ‘Others’

- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

For Do not call , do not email

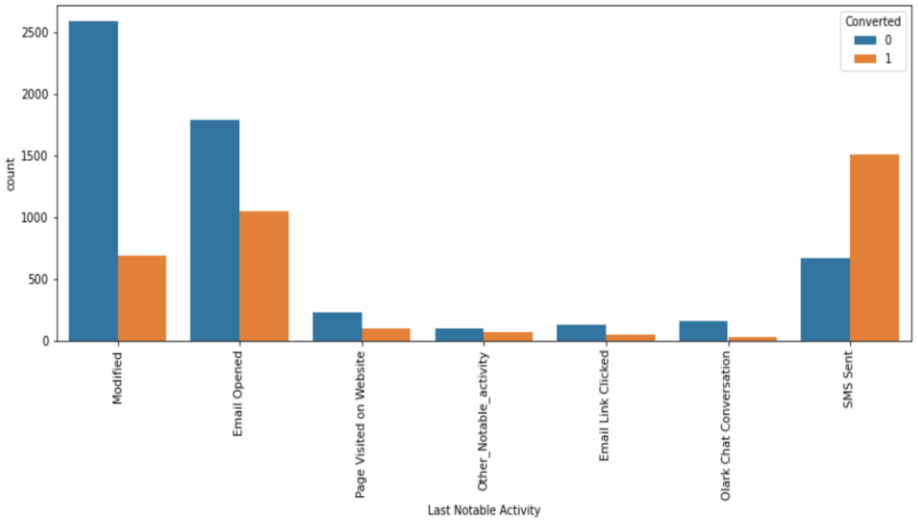
- We Can append the **Do Not Call** Column to the list of Columns to be Dropped since > 90% is of only one Value



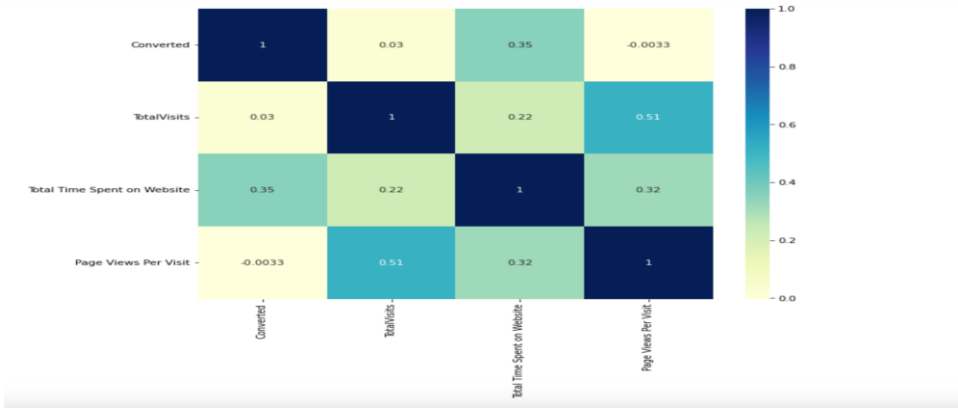


- list of columns we dropped due to low frequencies:

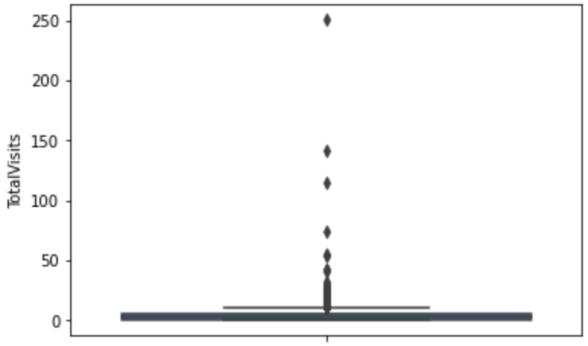
Country', 'What matters most to you in choosing a course', 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque']

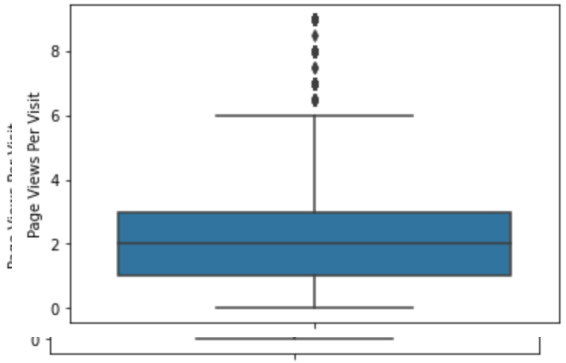


Checking correlations of numeric values

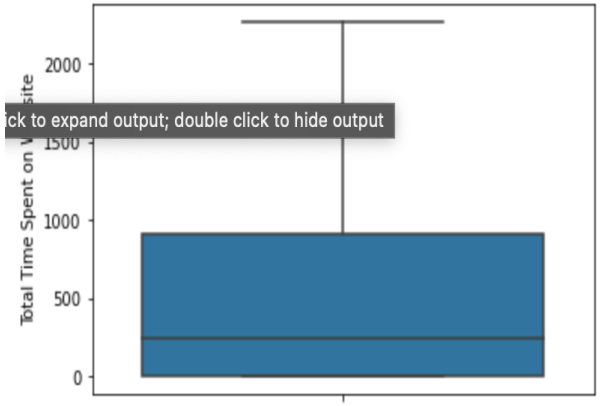


- Outlier treatments were done for :
- TotalVisits
- Page Views Per Visit
- Median for converted and not converted leads are the close.
- Nothng conclusive can be said on the basis of Total Visits





- No Outlier treatment was Done for
- Total Time Spent on Website’
- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.



- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion

MODEL BUILDING AND PREDICTION

- Dummy variables were created
- Train and test split was done on 70-30% and data was scalled
- Build models using Logistic regressions and rfe selections , reduced the features till p value was low and VIF values less than 5. we selected the third model with features as in image.

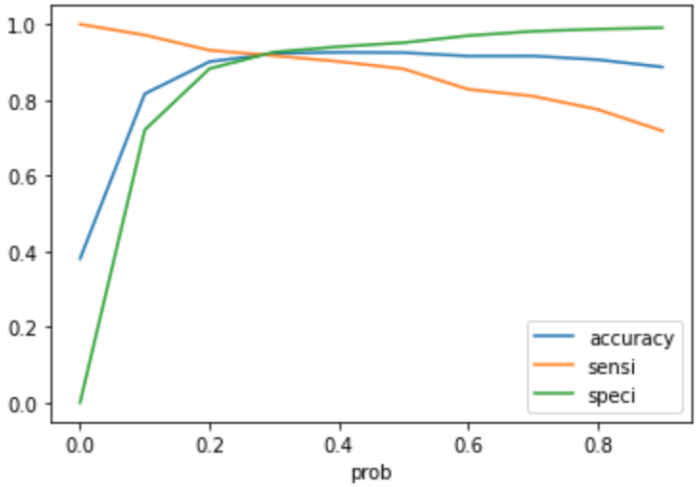
]:

	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

PLOTTING ROC CURVE

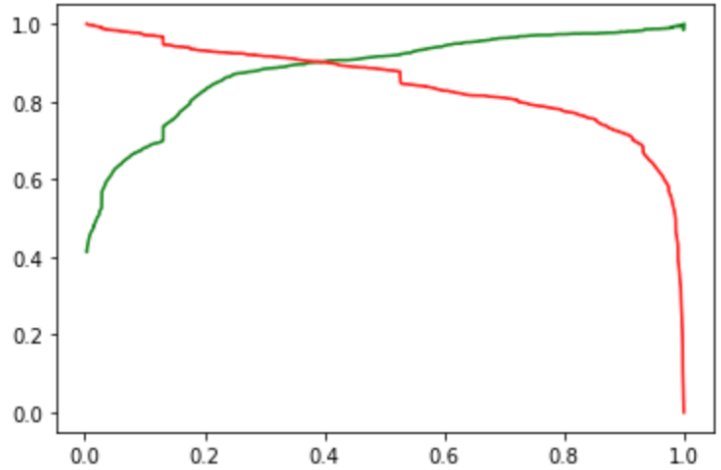
The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%



After running the model on the Test Data these are the figures we obtain:

- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%



FINAL OBSERVATION

- **Train Data:**
- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%

Test Data:

•Accuracy : 92.78%

•Sensitivity : 91.98%

•Specificity : 93.26%

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model