Lead Case Study

Neelima & Pratyusha

2nd March 2020

Business Goal

The business goal of the X education company is to improve the conversion rate of leads from the current 30% to 80%

Steps for building

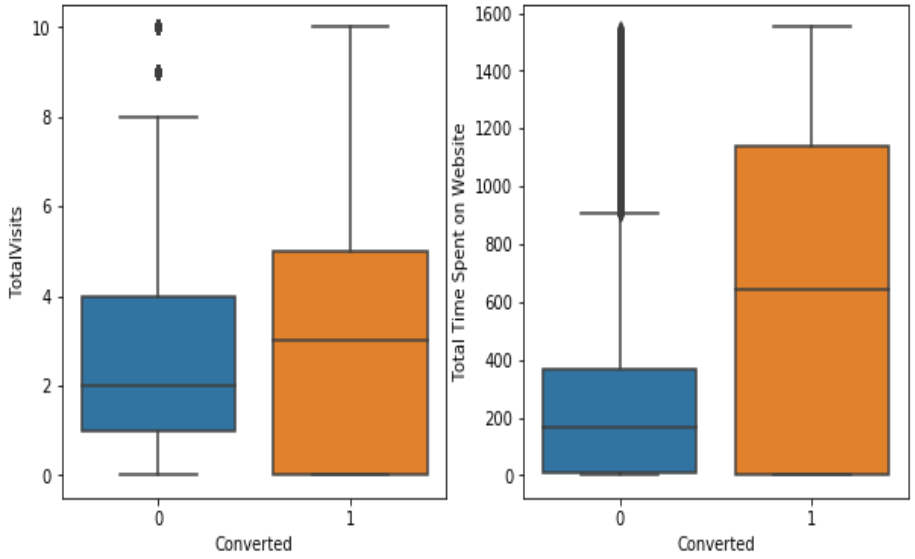
- Data Inspection
- Data Cleaning
- Data Preparation
- Data Visualization
- Dummy Variable Creation
- Split the data into the Train and Test set
- Scaling
- Feature selection by RFE
- Logistic Regression Modeling
- Prediction on Test Set
- Assigning Lead score to the dataset

Data Cleaning

Features	% of Null values
Lead Number	0
Lead Origin	0
Lead Source	0.39
Do Not Email	0
Do Not Call	0
Converted	0
TotalVisits	1.48
Total Time Spent on Website	0
Page Views Per Visit	1.48
Last Activity	1.11
Country	26.63
Specialization	15.56
How did you hear about X Education	23.89
What is your current occupation	29.11
What matters most to you in choosing a course	29.32
Search	0
Newspaper Article	0
X Education Forums	0
Newspaper	0
Digital Advertisement	0
Through Recommendations	0
Tags	36.29
Lead Quality	51.59
Lead Profile	29.32
City	15.37
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Score	45.65
A free copy of Mastering The Interview	0
Last Notable Activity	0

- Features which are having more than 40% of null values are dropped
- Features which had single category which would not give any value add in analysis are also dropped

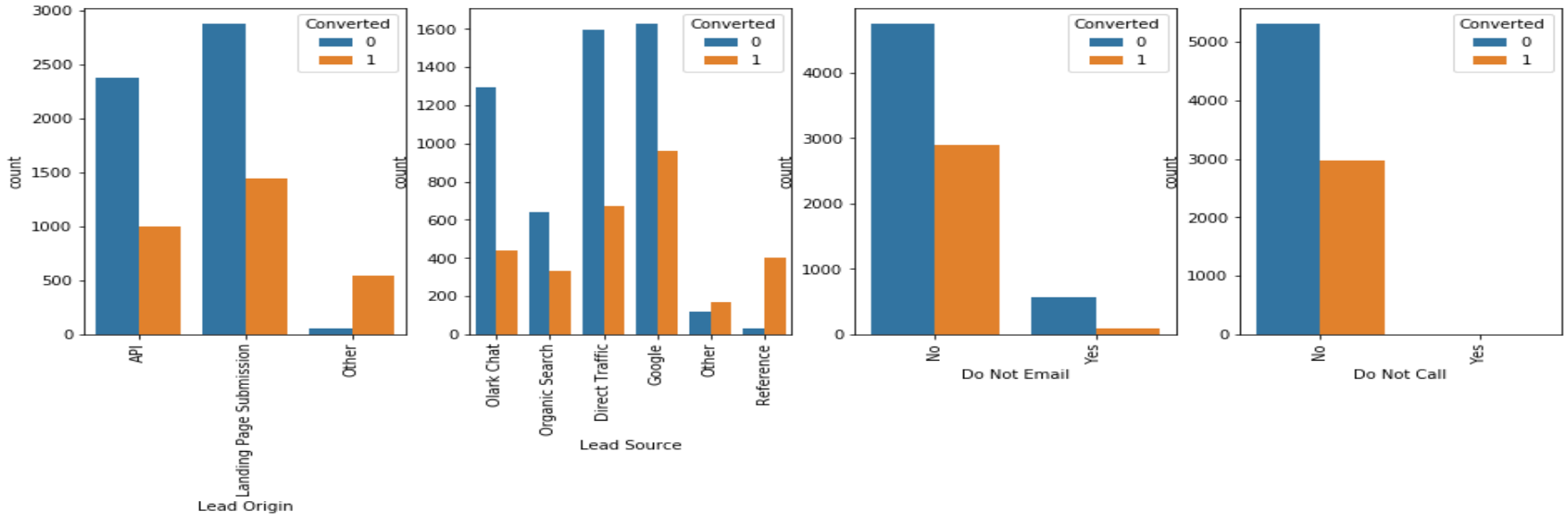
Data Visualisation



- The number of conversions are more for people who have visited the website more number of times
- The number of conversions are more for the people who have spent more time on the website

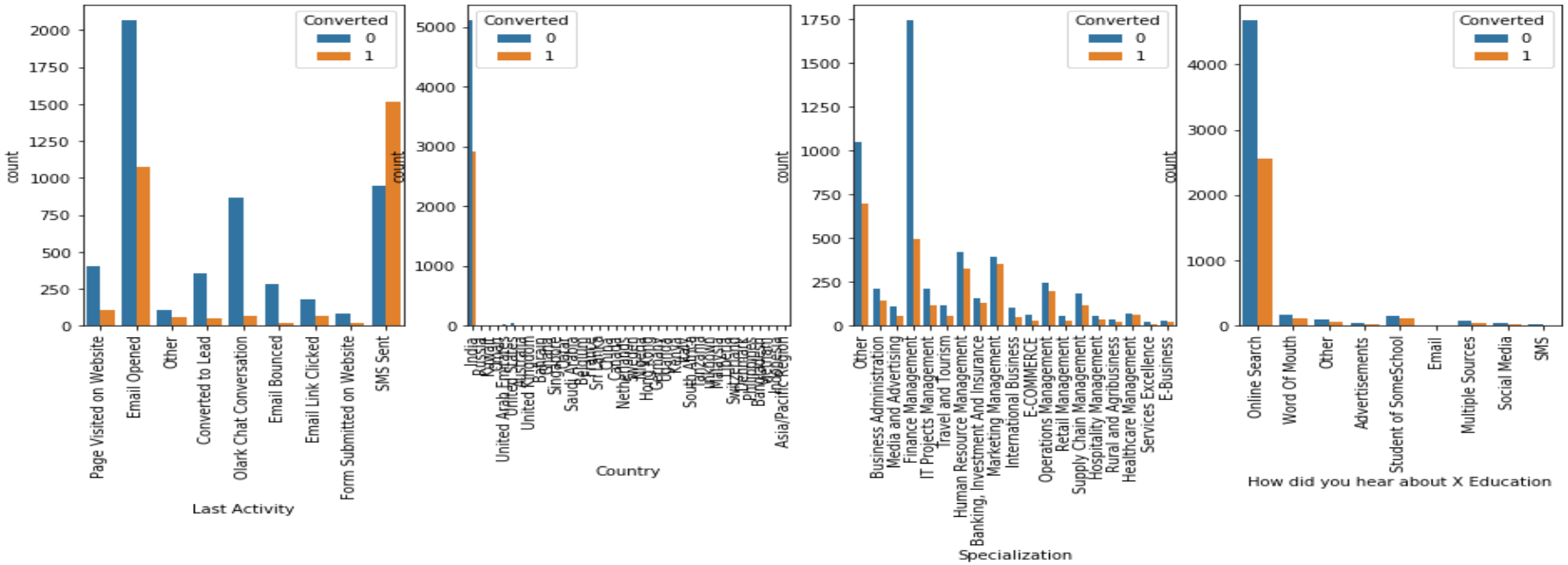
- Total visits and Page views per visit are highly correlated so one of the feature is dropped

Data Visualization



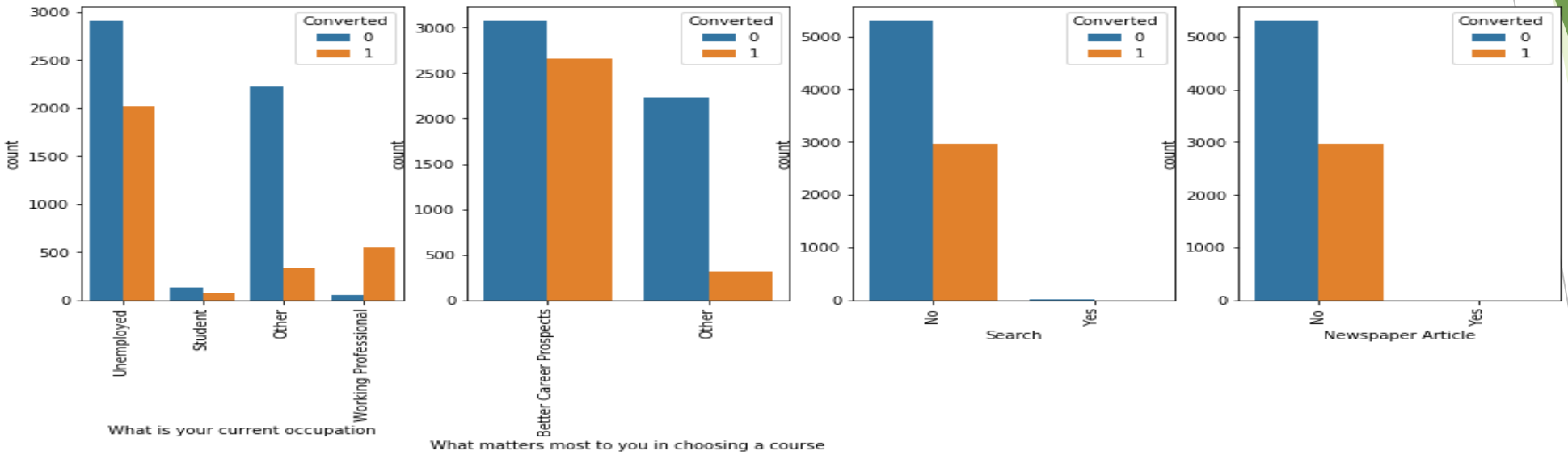
- Landing Page Submission category has more number of conversions when compared to other categories
- For category ' Google' has more number of hot leads
- Do Not Email : ‘No’ category has more number of hot leads
- Do Not Call : ‘No’ category has more number of leads.

Data Visualisation



- ‘Last Activity’ feature’s SMS sent category has more number of hot leads
- ‘Country’ feature is not giving much information and so it is dropped
- ‘Specialization’ feature’s Other category has more number of hot leads
- ‘How did you Hear about X education’ feature’s online search has more number of hot leads

Data Visualisation



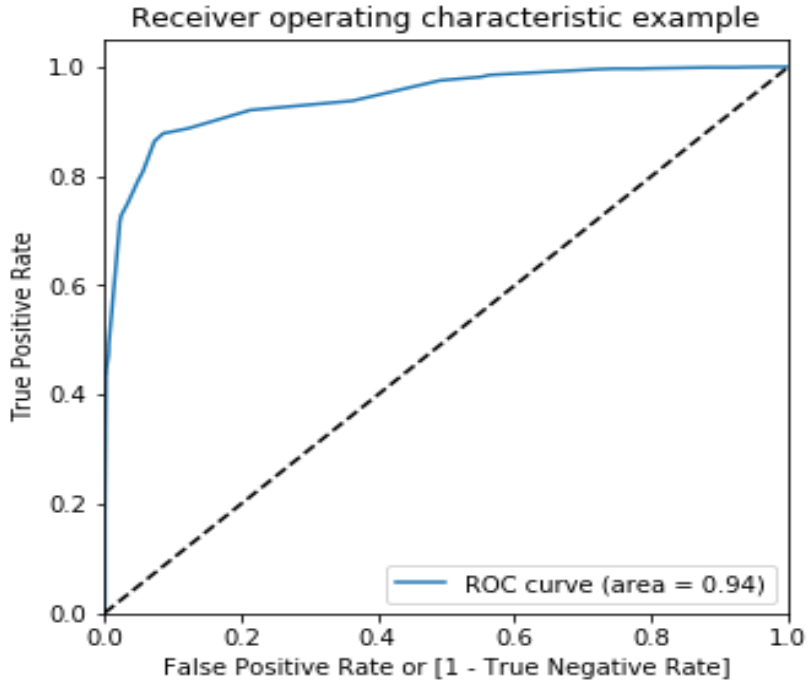
- ‘What is your current occupation’ feature’s Unemployed category has more number of hot leads
- ‘What matters most to you in choosing a course’ feature’s Better career prospects has more number of hot leads
- ‘Search’ and ‘Newspaper Article’ features are not giving much information, so these features are dropped

Data Preparation & Scaling

- Categories with less number of values in a feature are combined to a new category ‘Other’
- Outliers are treated with IQR
- Categorical variables are handled by creating dummy variables
- Standardised scaling is done for the numerical columns
- Data is divided into Train and Test in 70:30 ratio

Logistic Regression

- Logistic regression is applied on the features selected by RFE
- Features with less than 0.05 P value are dropped
- VIF values for all the features are maintained below 5



- ROC curve has been created with False Positive Rate on X-Axis and True Positive Rate on Y- Axis
- The area under the curve is 94% indicating the high accuracy of the model

Logistic Regression model performance on Train and Test datasets

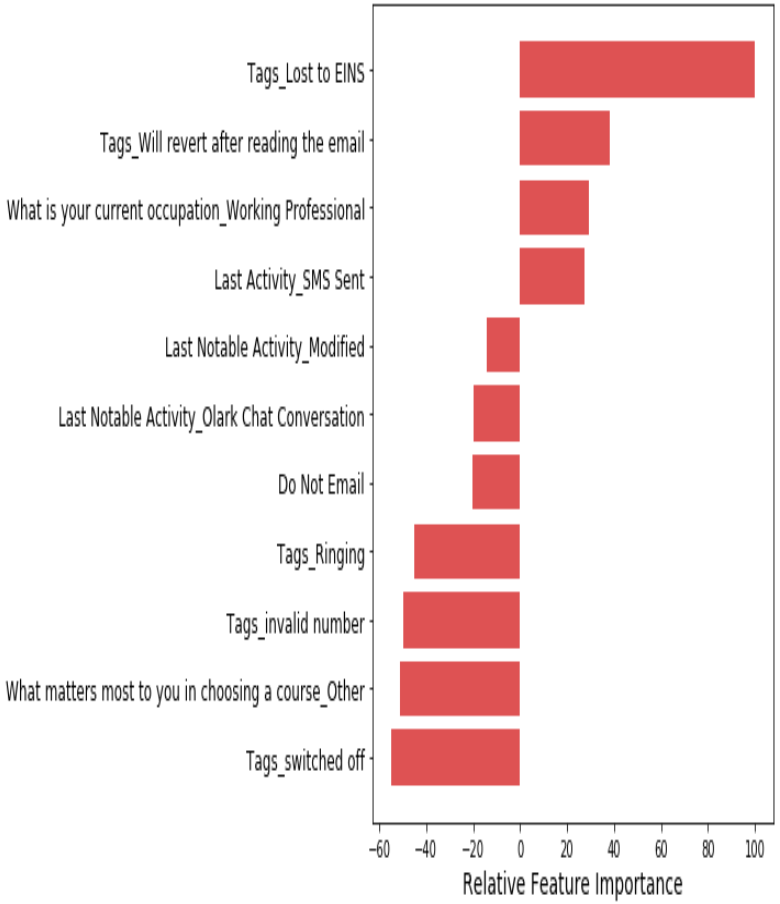
Train Set

- Accuracy: 90%
- Sensitivity: 88%
- Specificity: 91%
- False Positive Rate: 8.7%
- Positive Predictive Value: 85%
- Negative Predictive Value: 92%
- Precision: 85%
- Recall: 88%

Test Set

- Accuracy: 89%
- Sensitivity: 87%
- Specificity: 89%
- False Positive Rate: 10%
- Positive Predictive Value: 81%
- Negative Predictive Value: 92%
- Precision: 81%
- Recall: 87%

Feature importance



The conversion probability will increase with increase of these values:

- Tags_Lost to EINS
- Tags_Will revert after reading the email
- What is you current occupation_Working Professional
- Last Activity_SMS Sent

The conversion probability decreases with decrease of these values:

- Last Notable Activity_Modified
- Last Notable Activity_Olark Chat
- Do Not Email
- Tags_Ringing
- Tags_Invalid Number
- What matters most to you in choosing a course_Other
- Tags_switched off

Final Dataset with conversion Probability and Lead Score

LeadID	Converted	Converted_Prob	Final_Predicted	Lead_Score	Lead Number
2764	1	0.999953	1	100	633120
3519	1	0.999953	1	100	626813
5784	1	0.999953	1	100	605335
5806	1	0.999953	1	100	605266
6586	1	0.999953	1	100	599270
3829	1	0.999869	1	100	623382
6579	1	0.999869	1	100	599326
8867	1	0.999869	1	100	582296
3192	1	0.999597	1	100	629451
3288	1	0.999597	1	100	628500
7853	1	0.999597	1	100	589544
8117	1	0.999597	1	100	587883
746	0	0.999534	1	100	652708
2127	1	0.999534	1	100	639297
2354	1	0.999534	1	100	637070
2475	1	0.999534	1	100	635910
2725	1	0.999534	1	100	633515
3006	1	0.999534	1	100	630972
3185	1	0.999534	1	100	629511
3751	1	0.999534	1	100	624227
3790	1	0.999534	1	100	623770
4312	1	0.999534	1	100	618435

The final dataset after building logistic regression model with conversion probability and Lead_Score

Thank You ☺

-Neelima & Pratyusha