Lead Scoring Case Study

By

- Amber Kaushal

Business Scenario

- An education company named X Education sells online courses to industry professionals.
- 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.
- 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%.

Problem Statement

- X Education gets a lot of leads, its lead conversion rate is very poor
- For example, 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.

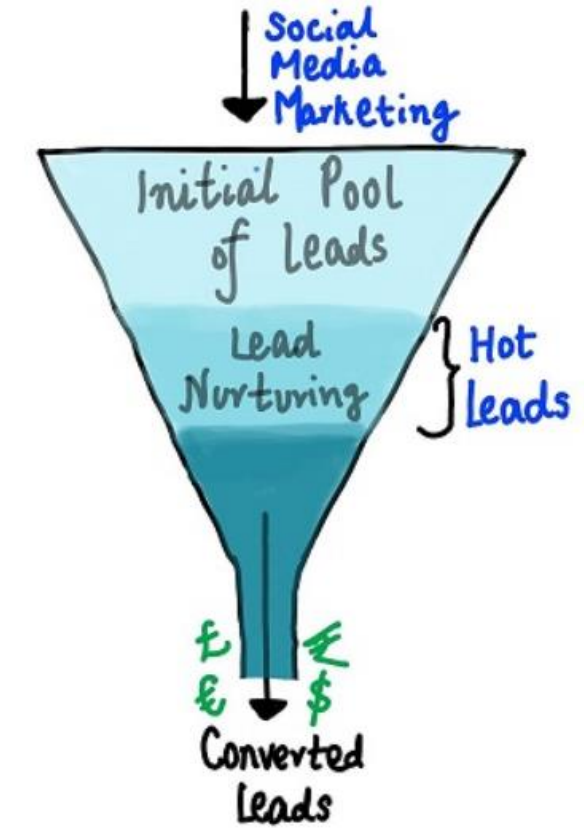
Case Study Objective

- We need to build a model to identify relevant leads which will most likely to convert.
- Model assign a lead score 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 CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Contents

- Solution Approach
- Analytical Approach
- EDA Plots Visualization
  - Categorical Variables
  - Numerical Variables
- ROC & Precision Curve
- Performance of Final Model
- Inferences / conclusion
  - Recommendations

Solution Approach



Identify Hot Leads

We will identify the leads which are most probably to convert

Focus On Hot Leads

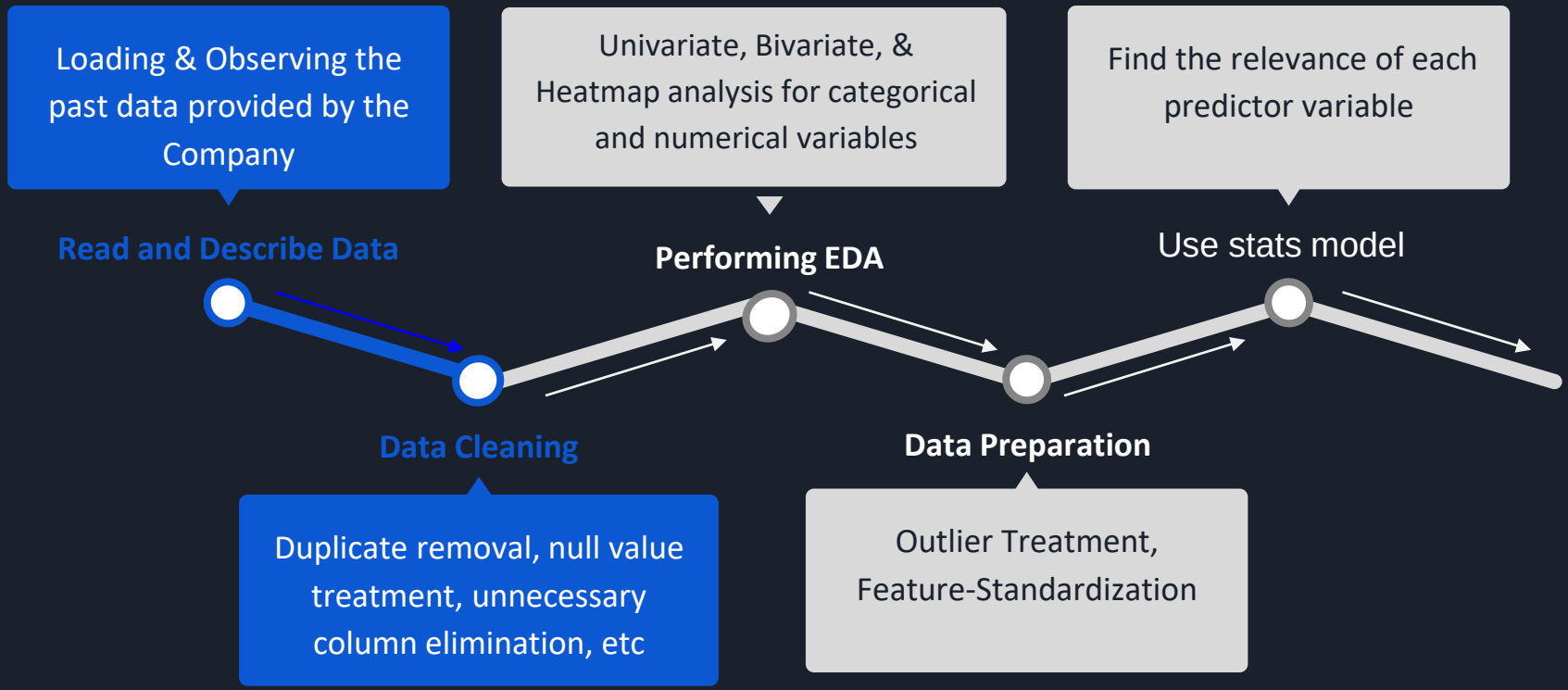
Now we will target the smaller set of leads with effective communication and try to convert them.

Conversion

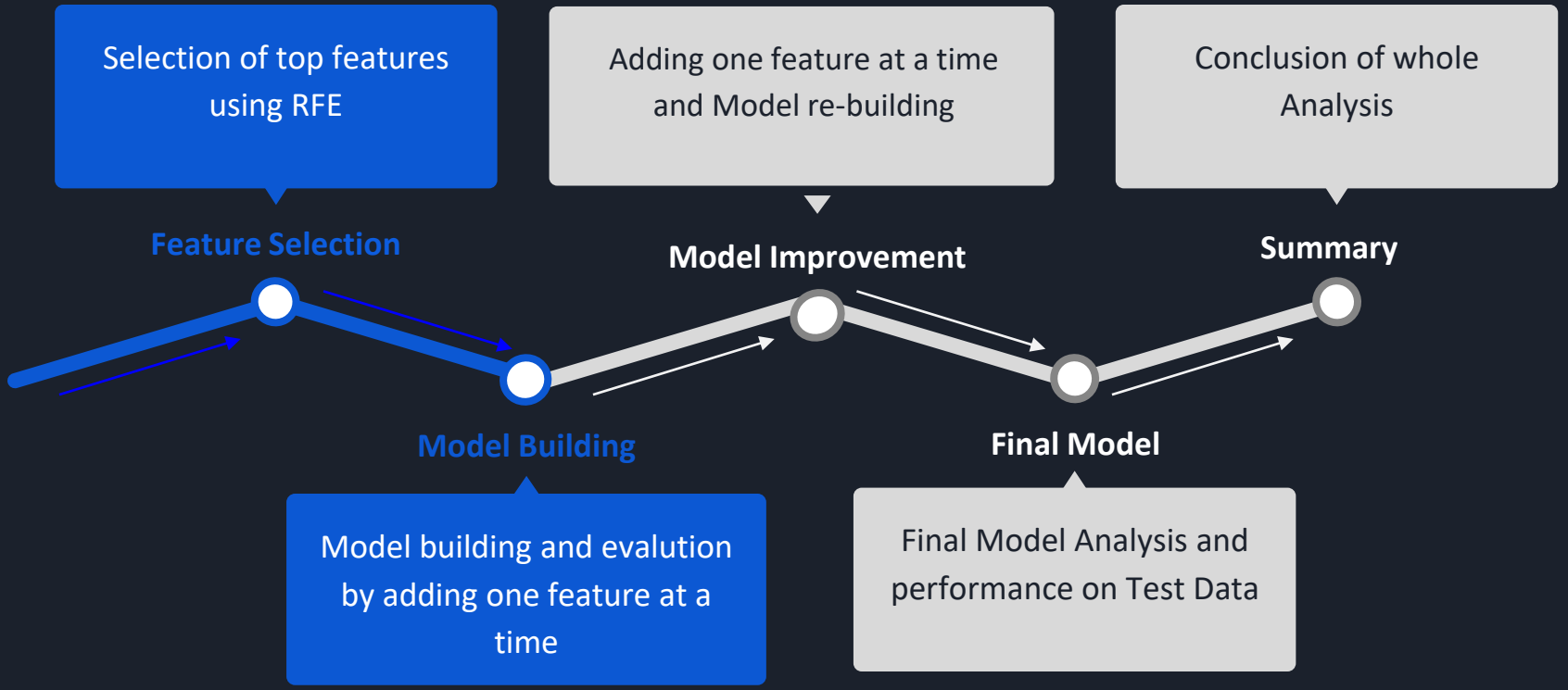
Since we more focused on hot leads which were promising, we would have a better conversions and we can achieve an 80% conversion rate

Analytical Approach

Analytical Approach

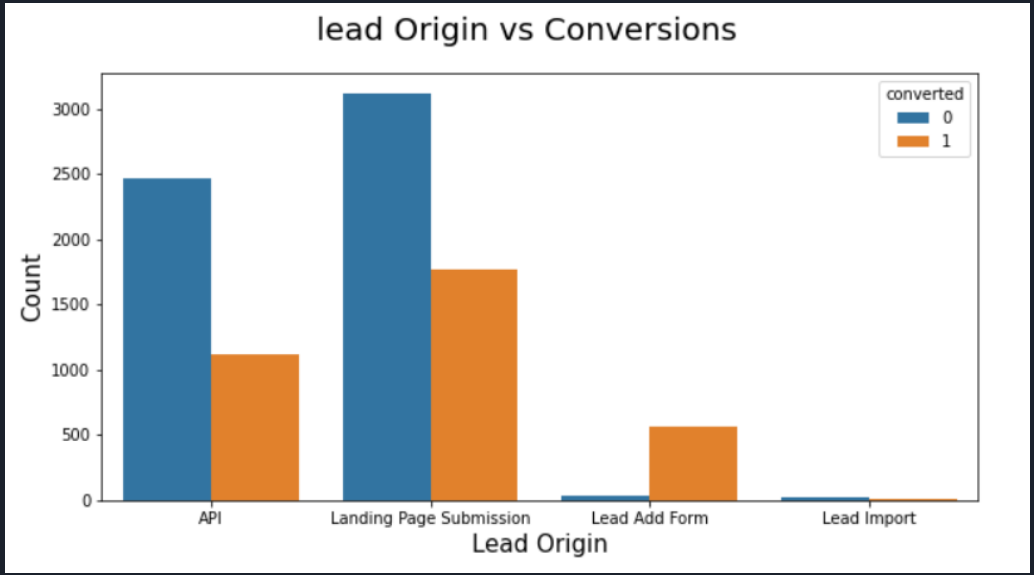


Analytical Approach



EDA Plots Visualization

Categorical Variables

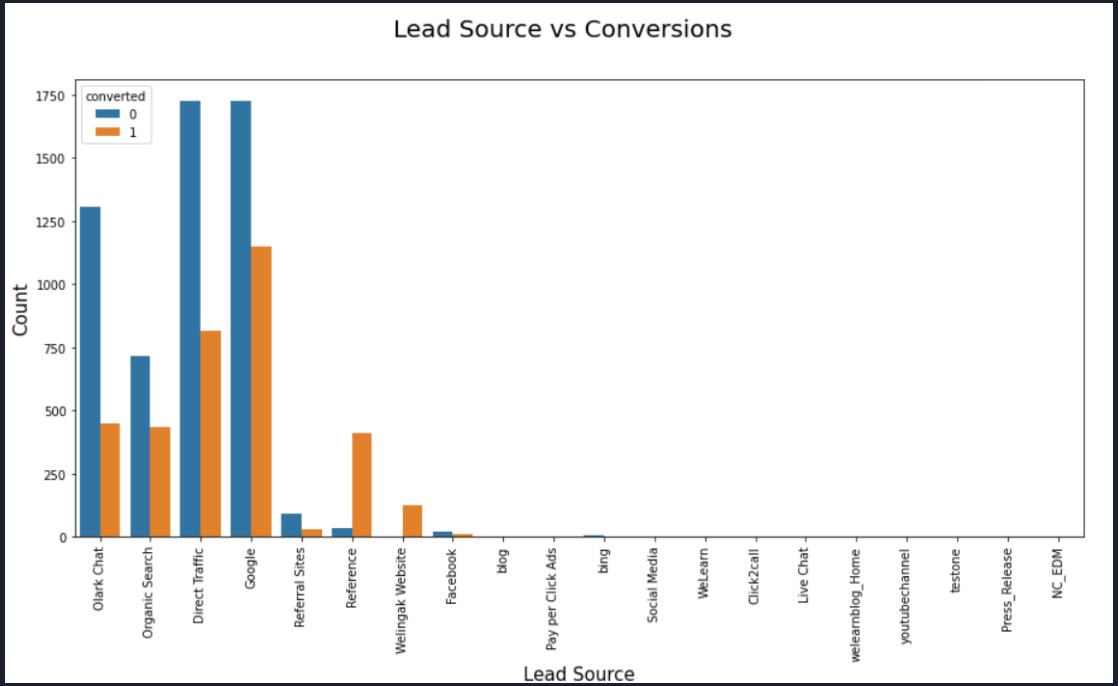


EDA plot depicting variation in categorical column "Lead Origin"

Inferences -

- Landing Page Submission & API getting the higher number of leads. And the conversion rate is also impressive.
- Lead Add Form has a higher conversion rate but the number of leads is less.
- Landing Page Submission & API has a significant number of leads, but we need to focus on the improvement of the conversion rate.
- The conversion rate of the Lead Add Form is higher but we need to generate more leads through Lead Add Form.

Categorical Variables

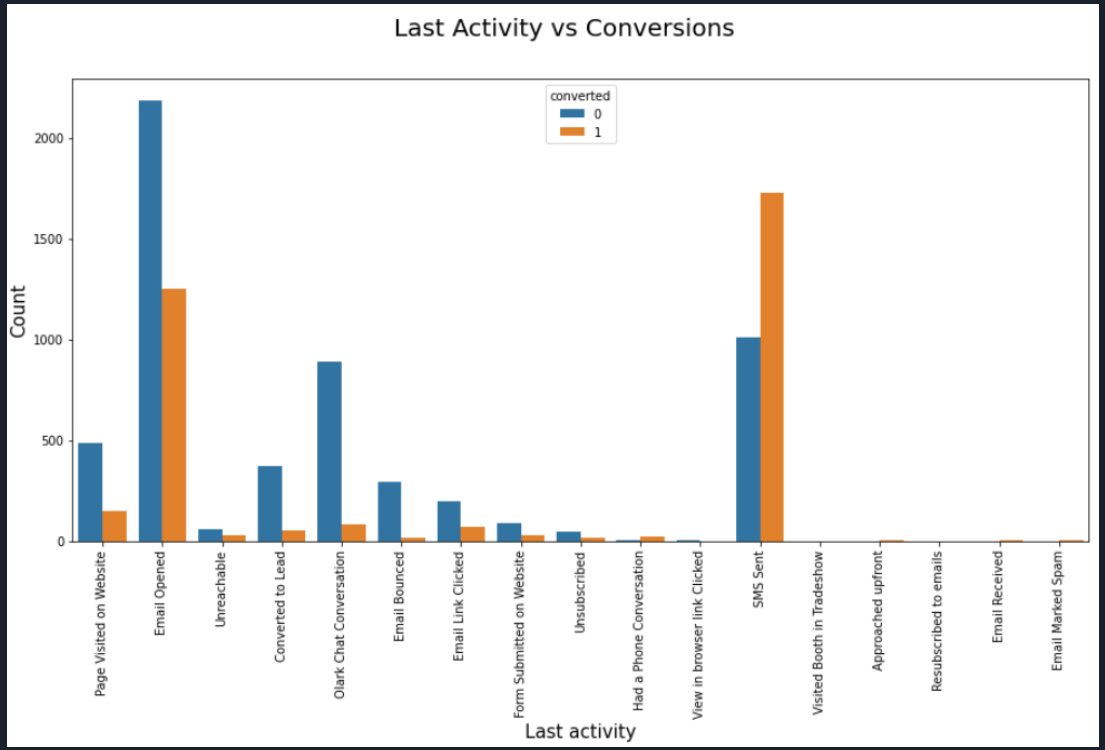


EDA plot depicting variation in categorical column "Lead Source"

Inferences -

- Google and Direct traffic generated a higher number of leads as well as conversions.
- "Reference" and "welingak Website" have a higher conversion rate but fewer leads.
- Olark Chat, Organic Search, Direct Traffic, and Google for these sources we need to focus on improvement of Conversion rate and need generate more leads from Reference and Welingak Website.

Categorical Variables

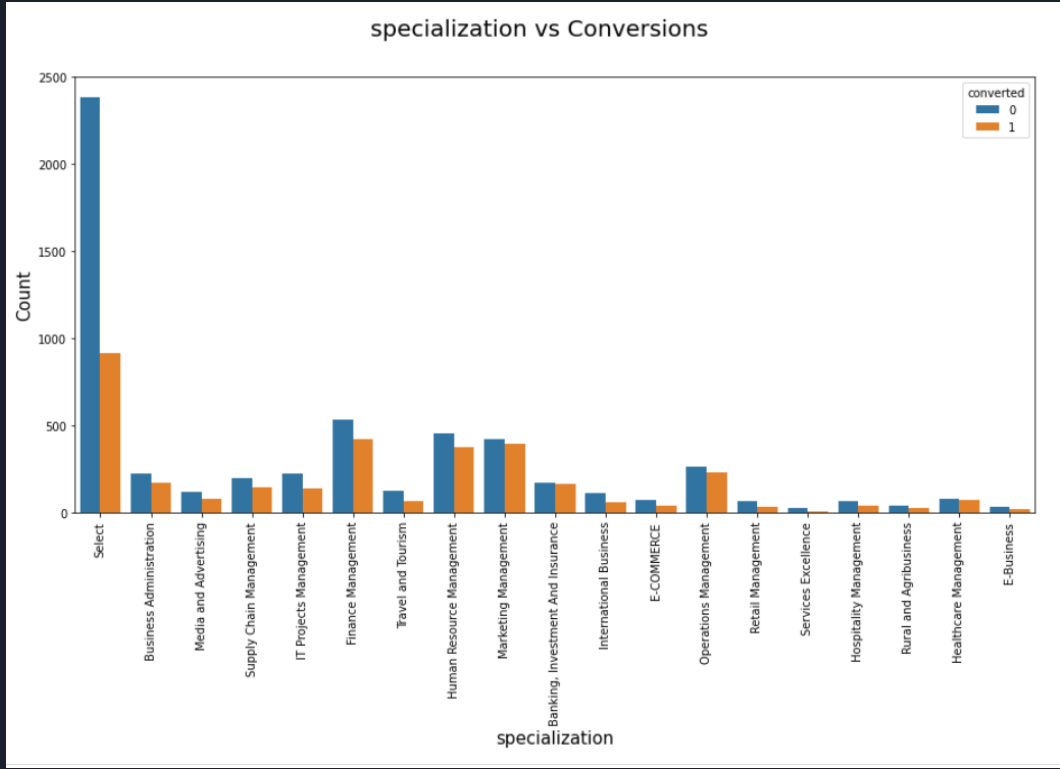


EDA plot depicting variation in categorical column "Last Activity"

Inferences -

- From the graph, we can see that the Last actives of the user "Email Opend" and "SMS sent" have a higher conversion rate.
- "Olark chat conversion" and "Page visited on Website" received a significant number of leads but the conversion rate is not so impressive.We need to try to convert them.

Categorical Variables

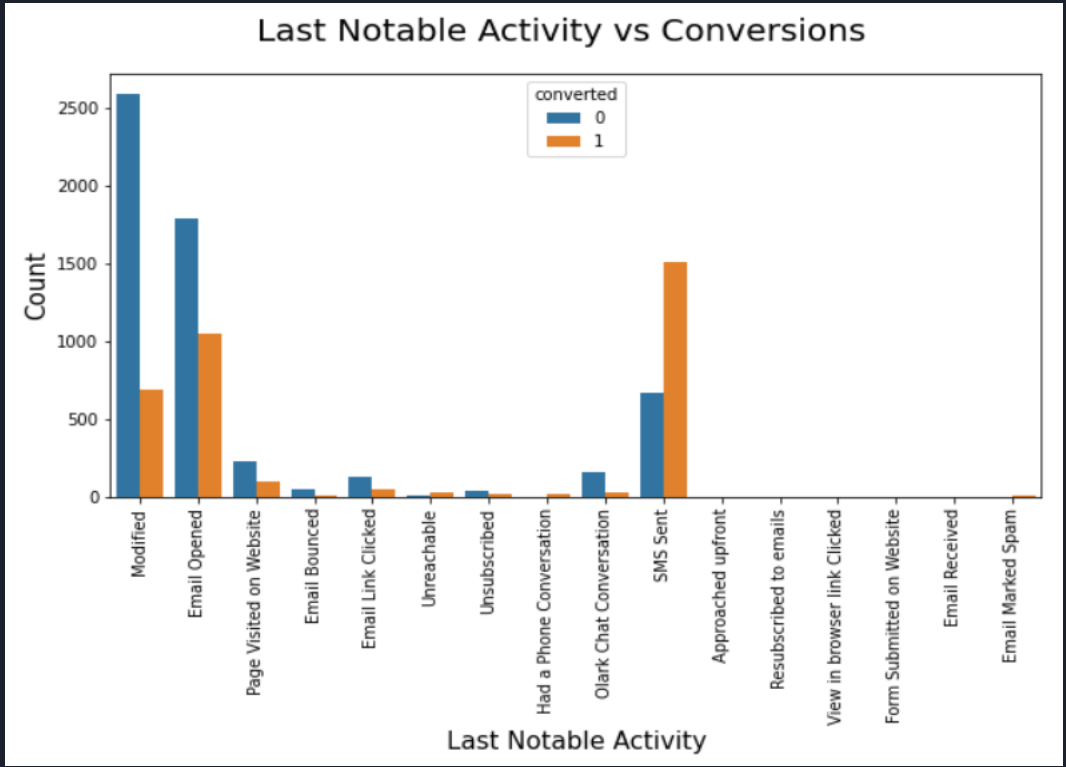


EDA plot depicting variation in categorical column "Specialization"

Inferences -

- Apart from "Select" (Which is a default option) Finance Management, Human Resource Management, & Marketing Management have good conversion rate, focus should be more on them.

Categorical Variables

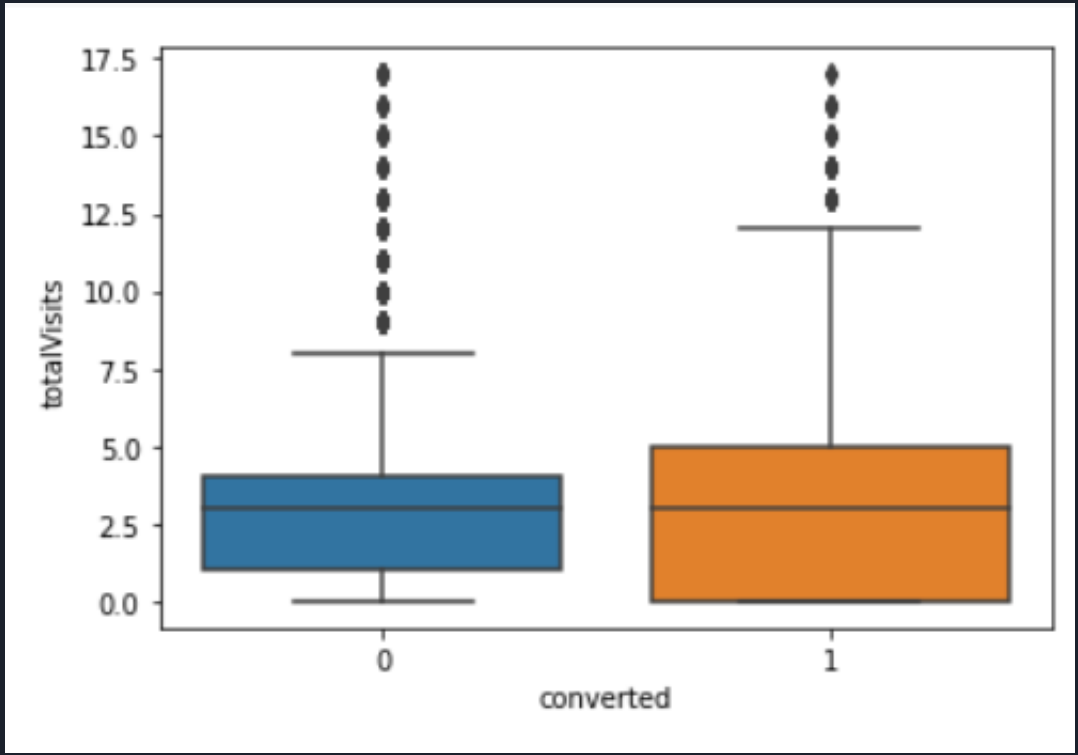


EDA plot depicting variation in categorical column "Last Notable Activity"

Inferences -

- As "Last Activity" column, We concluded the same for this column-From the above graph, we can see that the Last actives of the users "Email Opend" and "SMS sent" have a higher conversion rate.

Numerical Variables

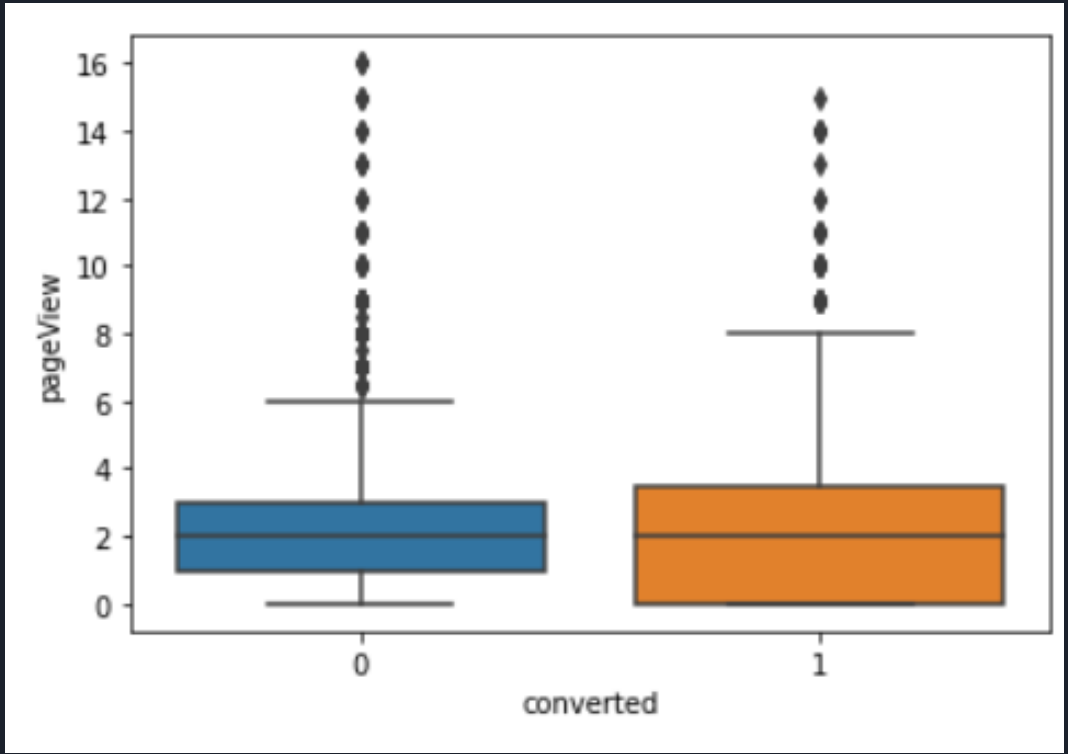


EDA plot depicting variation in numerical column "Total Visits"

Inferences -

- We have observed that the Median of both converted and not converted leads are seemed equal, We couldn't conclude on the basis of "Total Visits".

Numerical Variables

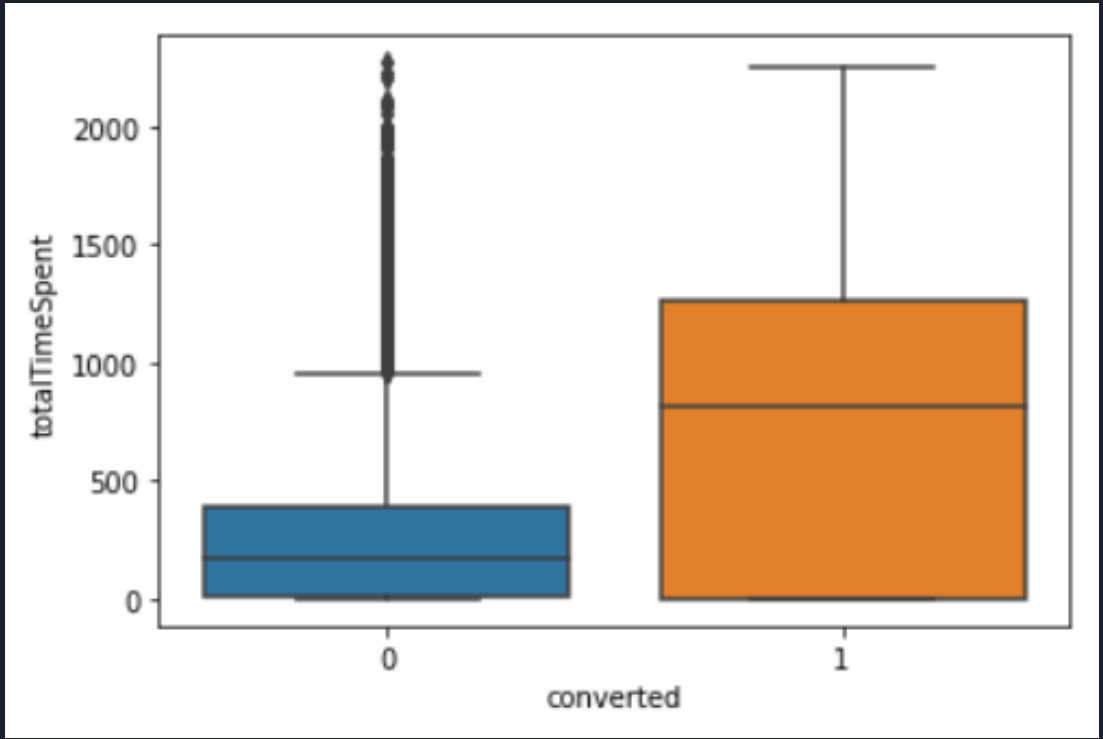


EDA plot depicting variation in numerical column "Page Views Per Visit"

Inferences -

- The median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit

Numerical Variables

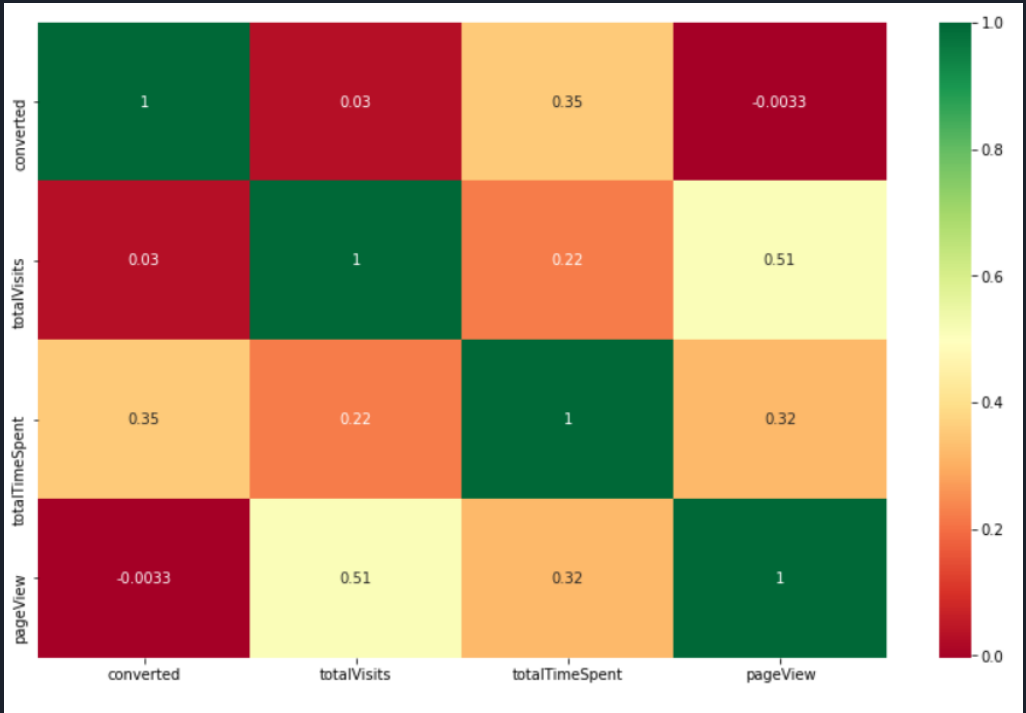


EDA plot depicting variation in numerical column "Total Time Spent"

Inferences -

- From the plot, it's clear that the person who spent more time on the website is more likely to convert.

Numerical Variables

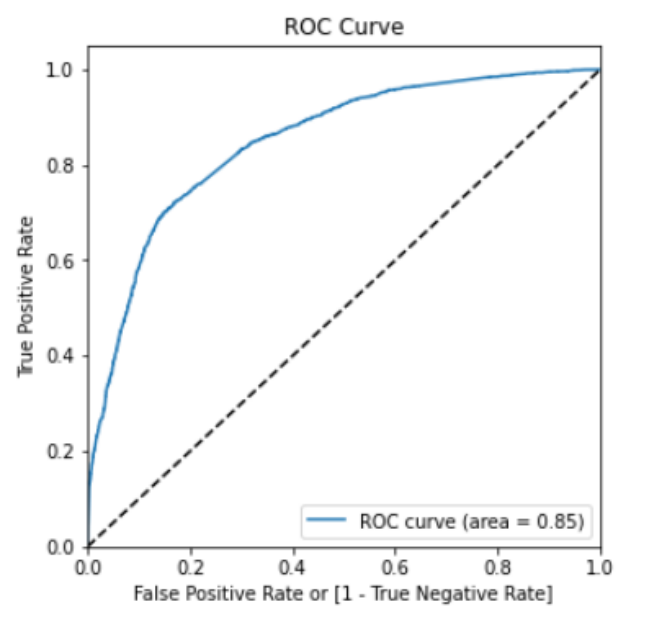


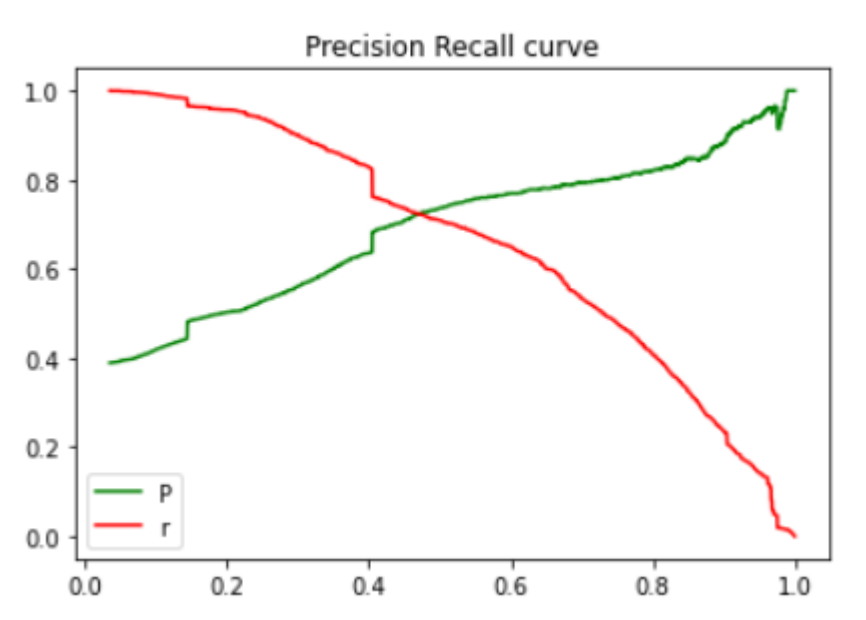
EDA plot depicting correlation of selected numerical columns

Observation -

- It is observed that "totalvisit" and "pageView" are highly correlated and have a correlation value of 0.51.

ROC & Precision Curve





Final Model Parameters:
Area under ROC = 0.85,

Threshold	Accuracy	Precision	Recall
0.1	0.47	0.42	0.99
0.2	**0.62**	**0.50**	**0.96**
0.3	0.69	0.56	0.90
0.4	0.75	0.64	0.83
0.5	0.79	0.74	0.71
0.6	0.79	0.77	0.65
0.7	0.77	0.79	0.53
0.8	0.74	0.82	0.41
0.9	**0.69**	**0.89**	**0.24**
1.0	0.62	0.00	0.00

We get the best **recall** at threshold value of **0.2**. It has good support of precision and accuracy.

We get the best **precision** at threshold value of **0.9**. It has good support of recall and accuracy.

Performance of Final Model

Final Model

Train Set
Accuracy Score: 0.79
Precision Score: 0.74
Recall Score: 0.71
F1 Score: 0.72

Test Set
Accuracy Score: 0.81
Precision Score: 0.75
Recall Score: 0.73
F1 Score: 0.74

Sensitivity: 0.70

Specificity: 0.84

Inferences / conclusion

Top Predictor variables

The important predictors in decreasing order of importance are as follows :-

- leadOrgn
- leadSrc
- totalTimeSpent
- specialization
- masteringInterview
- pageView
- matterMost
- Email

Many of these predictors are categorical. These are further explored to get the positive and negative categories.

Positive and Negative categorical variables.

Positive Impact Categories	Negative Impact Categories
NC_EDM Live Chat WeLearn Flexibility & Convenience Reference Olark Chat Welingak Website Click2call Social Media Google Facebook Direct Traffic Lead Add Form Healthcare Management Banking, Investment And Insurance Travel and Tourism E-COMMERCE Media and Advertising Business Administration Finance Management Organic Search masteringInterview Marketing Management	Lead Import International Business pageView bing Services Excellence API No selection Rural and Agribusiness Landing Page Submission Better Career Prospects Human Resource Management email Retail Management Other welearnblog_Home google youtubechannel testone blog Pay per Click Ads Press_Release

Recommendations

- The model is ready which will tell if the lead is hot lead or not. The sales team can use the model to find the hot leads.

- Depending upon the strategies, the sales team and choose the following threshold value
  - When the team has aggressive approach and wants to call all potential leads.
    - The team can choose the threshold value of **0.2** and get the hot leads from the model.
  - When the team has relaxed approach and wants to make only necessary call.
    - The team can choose the threshold value of **0.9** and get hot lead which has highest percentage of getting converted.

Score

- Each lead has been provided the score.
- Higher score value from 400 more is the chance of conversion
- Lower score value from 400 less is the chance of conversion