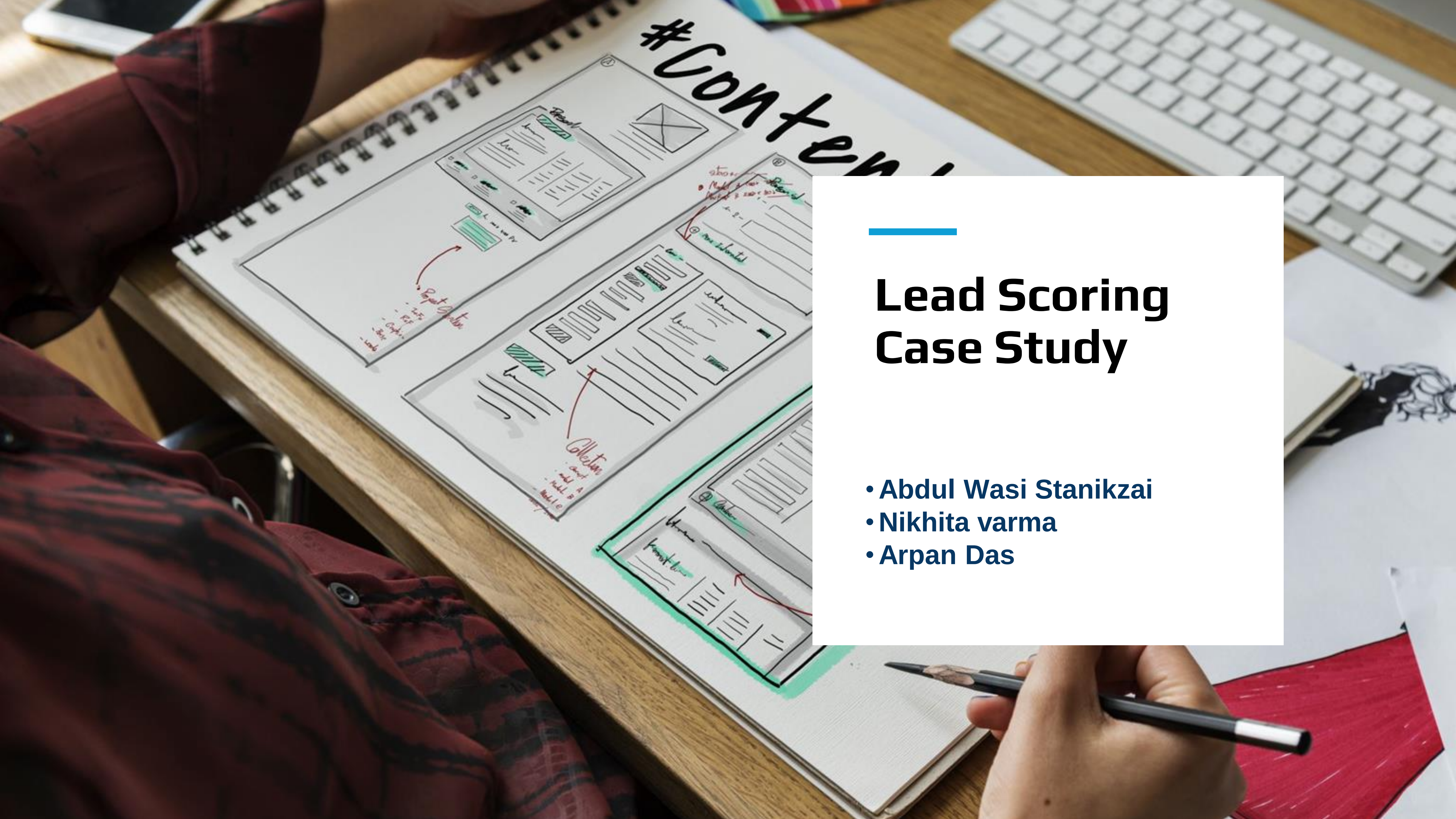
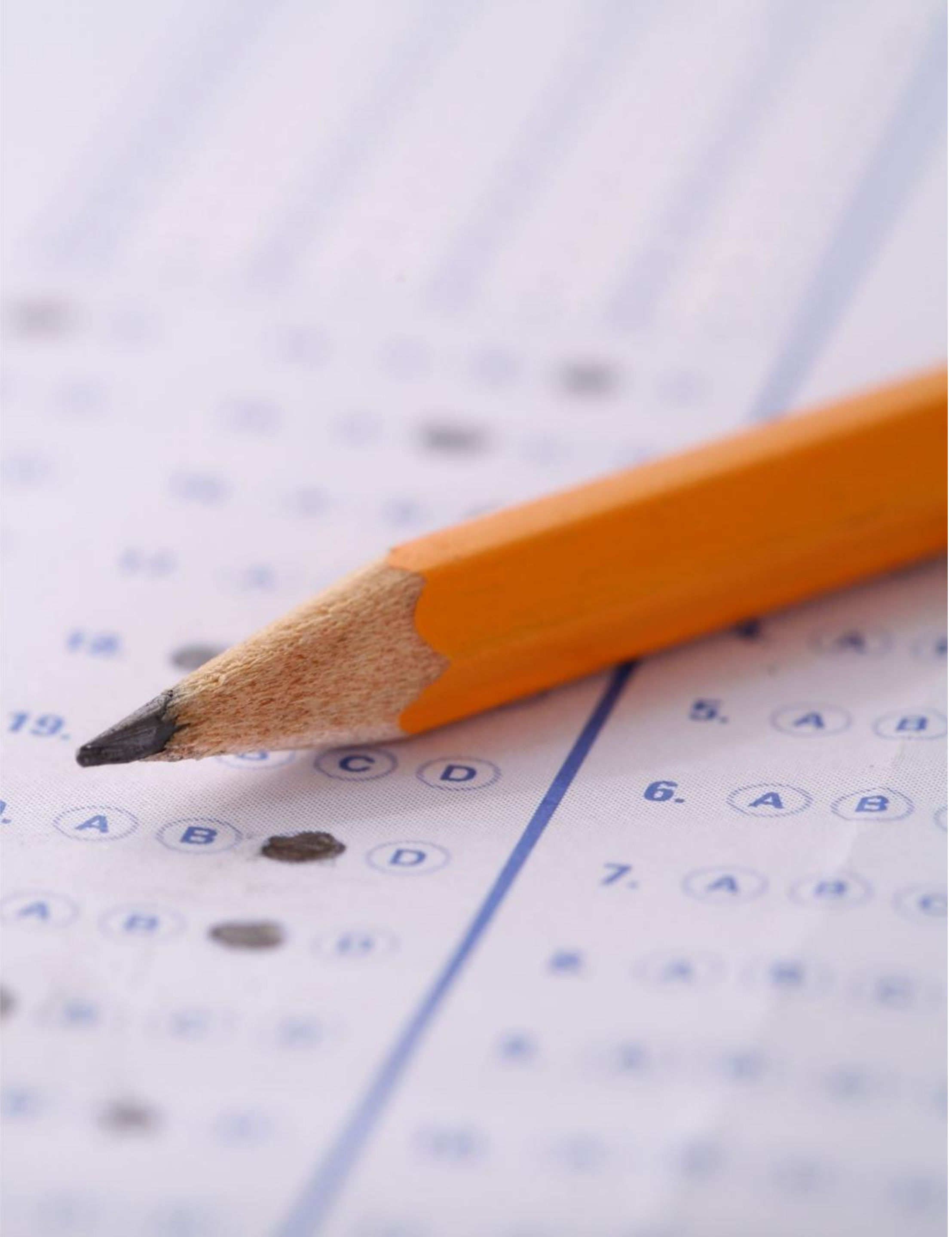


Lead Scoring Case Study

- Abdul Wasi Stanikzai
- Nikhita varma
- Arpan Das



Problem Statement

•X Education sells online courses to industry professionals.

•X Education gets a lot of leads, its lead conversion rate is very poor. If they acquire 100 leads in a day, only about 30 of them are converted.

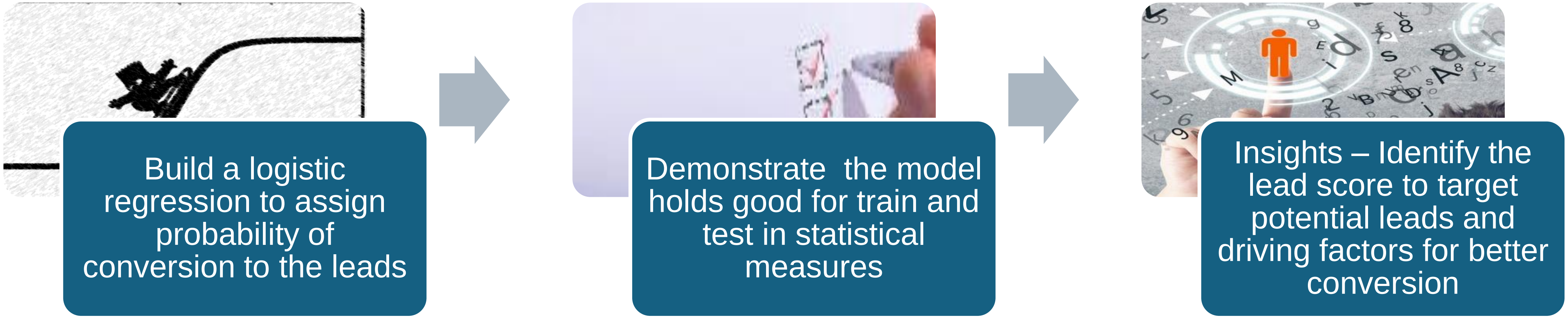
• X wants to Identify most promising leads known as 'Hot Leads'. i.e. the leads that are most likely to convert into paying customers so they can be targeted by Company' sales team for better conversion rate

- **Objective/ Goal :**

  Build a model to assign a lead score that will help the Company X target lead conversion rate of around 80%.

- **Data**
  - Leads dataset (based on form filled by leads) from the past with around 9000 data points
  - Leads dataset consists of various attributes deciding whether a lead will be converted or not.
  - The target variable, in this case, is the column 'Converted' - 0 means not converted and 1 implies converted
  - Data dictionary mentioning the details on the columns

Analysis/ Approach



Build a logistic regression to assign probability of conversion to the leads

Demonstrate the model holds good for train and test in statistical measures

Insights – Identify the lead score to target potential leads and driving factors for better conversion

Detailed Approach

1. Understand the Data
2. Data Preparation

i. Clean up of the data – select
ii. Handling missing values/ nulls
iii. Conversion of categorical variables (Yes/ No into 1/0, One-Hot Encoding for variables with more than two values
iv. Removing the columns with minimal / zero variance
v. Outlier handling

3. Explanatory Data Analysis
4. Build model with selected features (RFE). Eliminate features with high p values, multi collinearity
5. Final Model Evaluation using metrices
6. Decide the threshold to be used , predict in training data set
7. Test with test dataset
8. Final Conclusion – with lead score to be used and critical attributes

Data preparation and Feature Engineering

Categorical Variables

- Covert select to null as the option not selected by users
- Categorical value columns with only yes/no converted to 1/0

  ['Do Not Email', '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', 'A free copy of Mastering The Interview’]
- One hot encoding for Categorical values with more than two values

Deletions

- Drop the columns with more than 50% missing values

  'How did you hear about X Education', 'Lead Quality', 'Lead Profile’
- Drop columns that were scores and indexes with 45% missing values as we don’t know how thy are calculated so imputing them with mean/median for so many null values will tweak the data

  Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score,Asymmetrique Profile Score
- Drop city as 40% of values are null and there is no clear majority for the rest of values to impute – Mumbai 57% close to the missing values
- Drop the columns with 0 or minimal variance as they wont add any insights

  'Search','Magazine','Newspaper Article','X Education Forums','Newspaper','Digital Advertisement','Through Recommendations','Receive More Updates About Our Courses’, Update me on Supply Chain Content','I agree to pay the amount through cheque’,'What matters most to you in choosing a course’, Get updates on DM Content, Country
- Drop unique identifiers as they wont play a role in regression model

  Prospect ID', 'Lead Number'

Data preparation and Feature Engineering – continued

Imputations

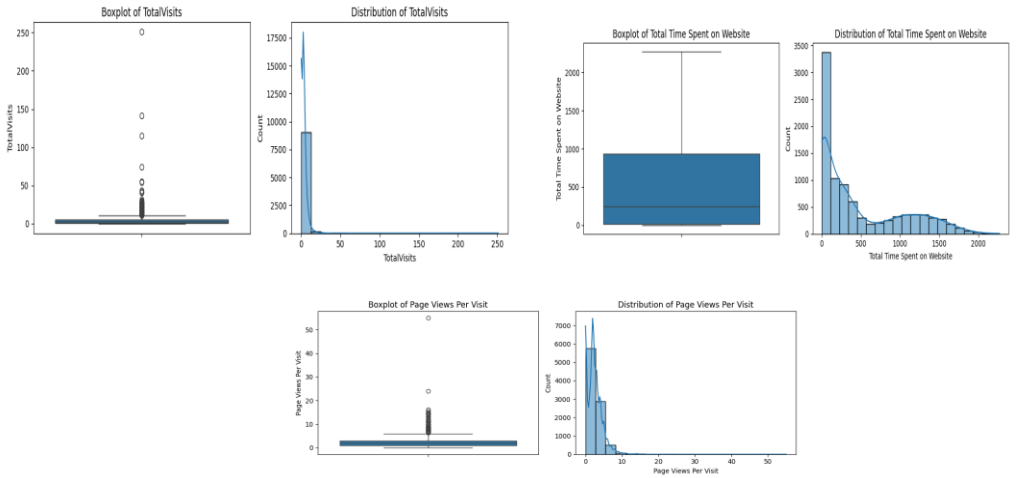
- Impute the columns with specific values based on maximum frequency or logic

```
1. Specialization – Null were marked as Others as rest of the specialization have been considered
2. Occupation - Null were marked as Unemployed ( max occurrence) as 85% of data is Unemployed
3. Tags – Null were  marked as 'Will revert after reading the email’ max occurrence
4. Last Activity had few nulls that were marked as ‘Email Opened’ max occurrence
5. The numerical columns TotalVisits and Page Views Per Visit have null values were imputed with Median
```

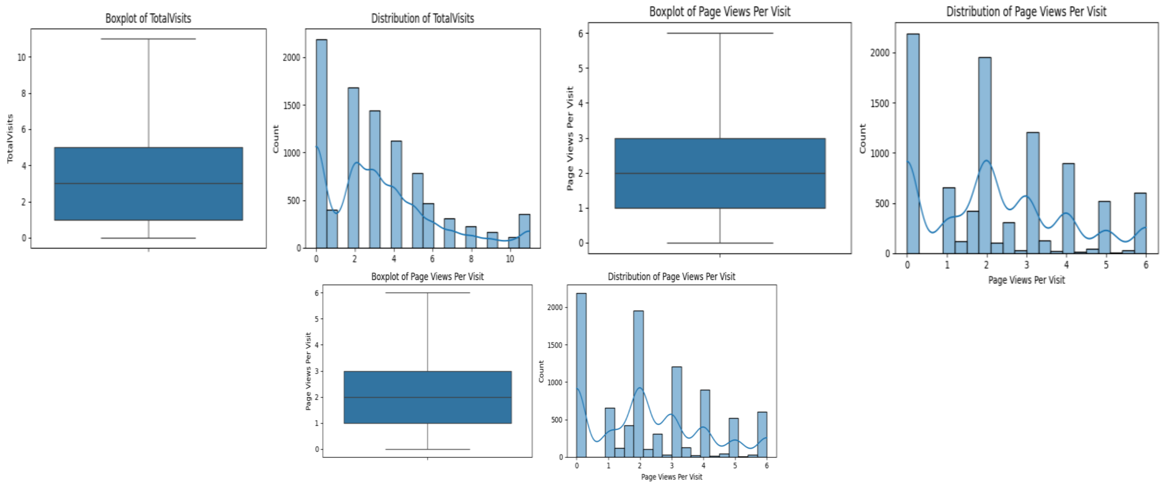
Outlier Handling

- Drop the columns with more than 50% missing values
- Numerical columns TotalVisits, ‘Total Time Spent on Website’, 'Page Views Per Visit’ had outliers
- For better regression – outliers were deleted based on IQR approach

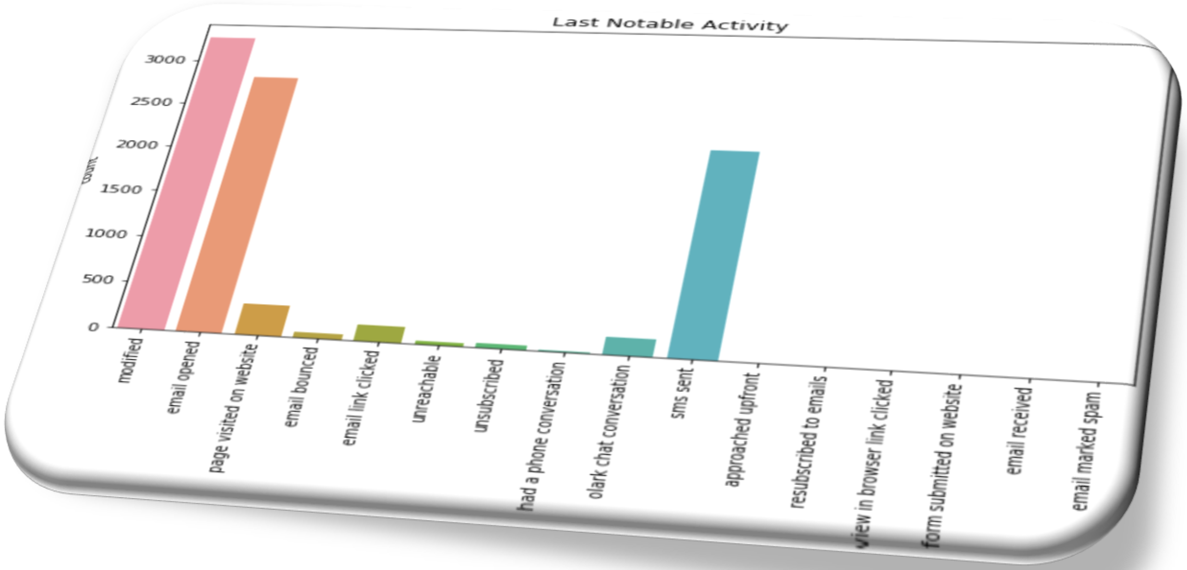
BEFORE

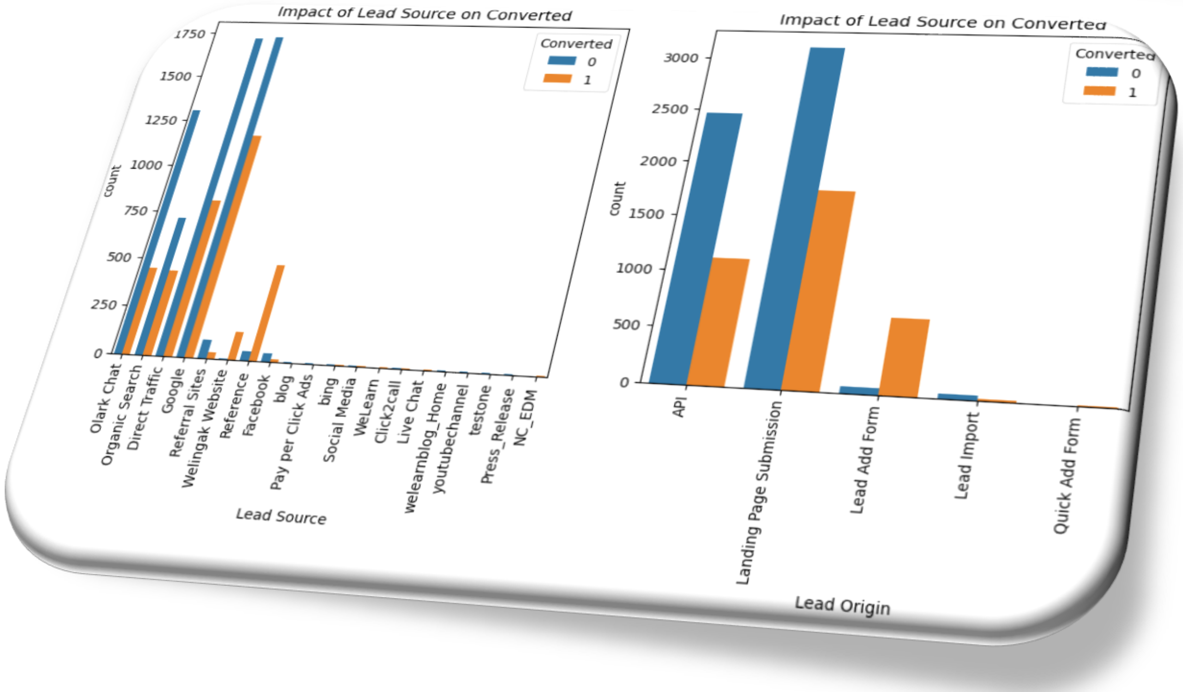


AFTER



EDA


Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam

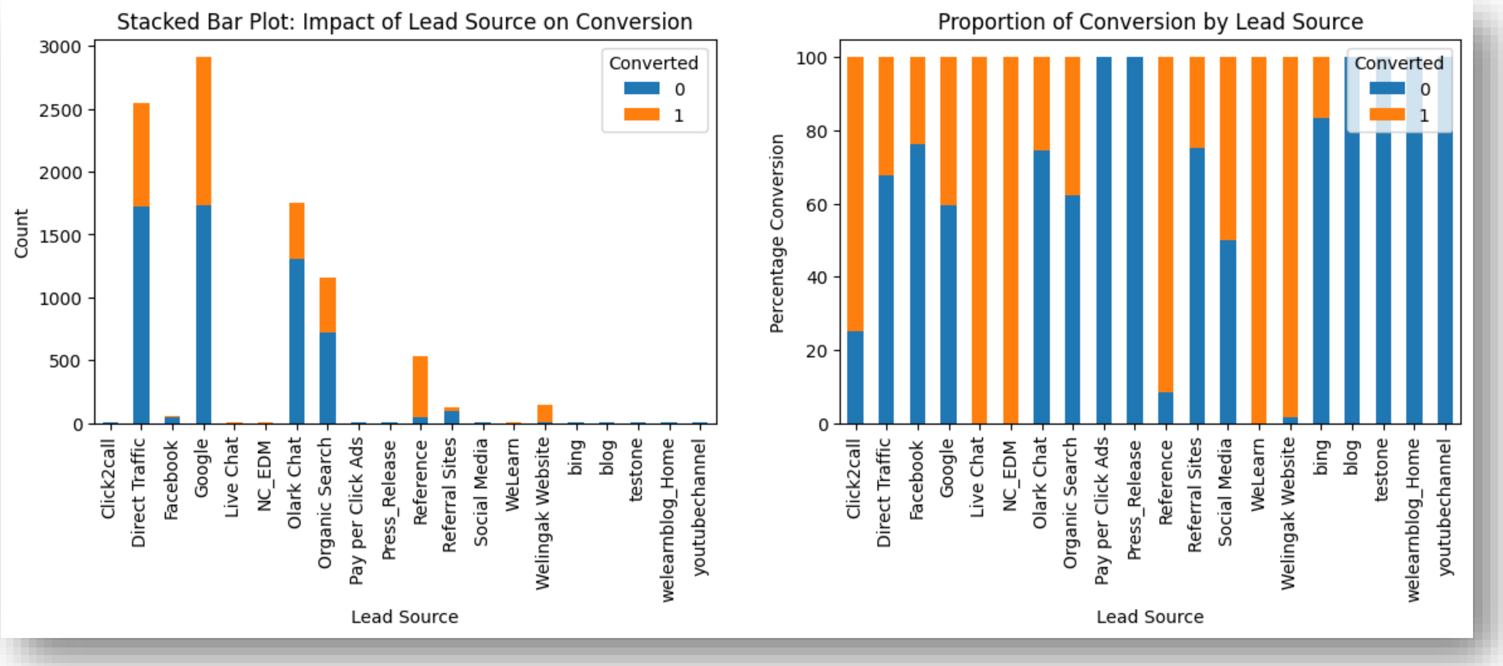

Impact of Lead Source on Converted
count
1750
1500
1250
1000
750
500
250
0
Converted
0
1
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source
Impact of Lead Source on Converted
count
3000
2500
2000
1500
1000
500
0
Converted
0
1
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin

Explanatory Data Analysis (1)

Categorical variables

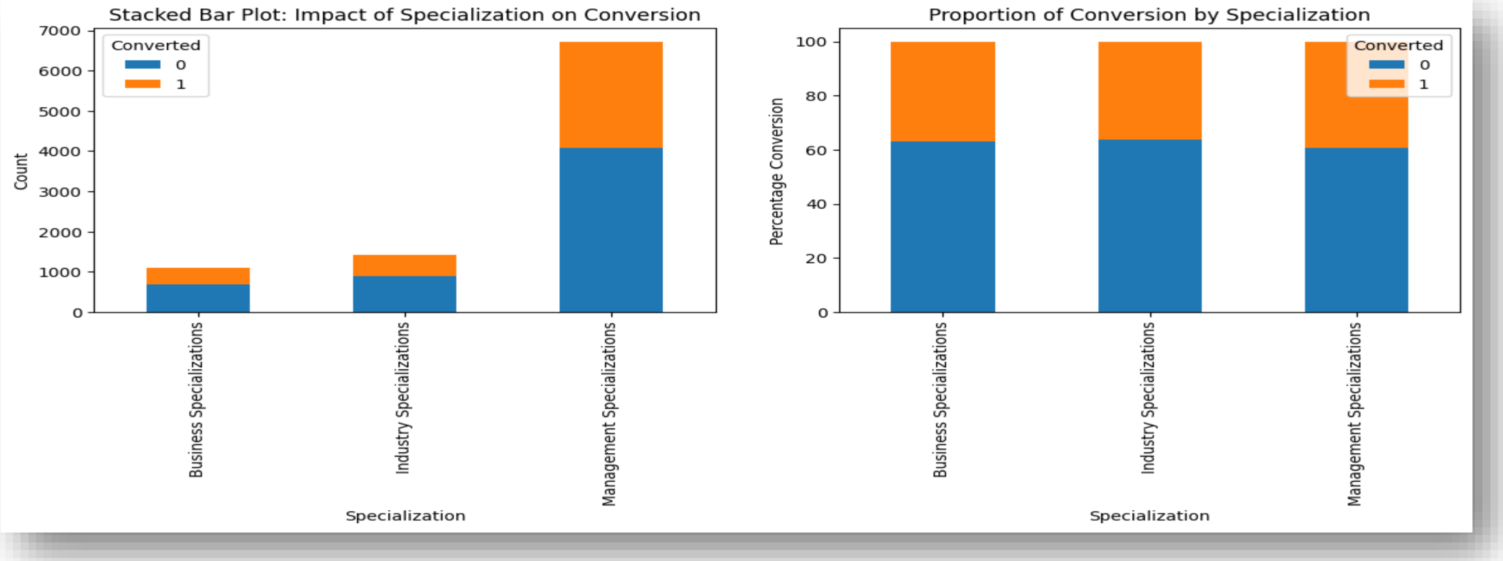
✓ Impact of Lead Source on Conversion

- While ‘Google’ and ‘Direct Traffic’ provide the maximum leads, ‘References’, ‘Welingak Website’ generate leads with decent number and best conversion rate
- ‘We learn’,’NC_EDM’,’Live Chat’ have 100% conversion, but their counts are lows



✓ Impact of Specialization on Conversion

- Management Specialization gets the maximum leads but there is no difference in the percentage of conversion based on the type of Specialization

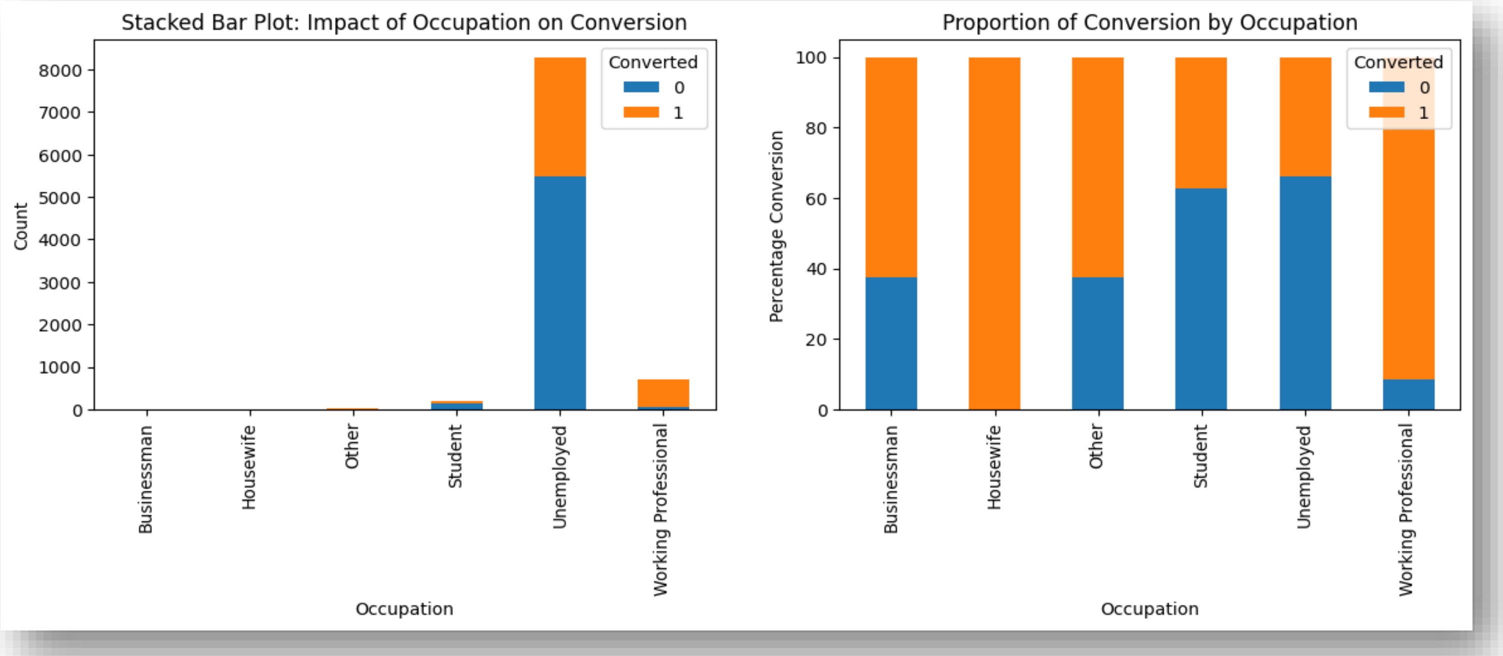


Explanatory Data Analysis (2)

Categorical variables

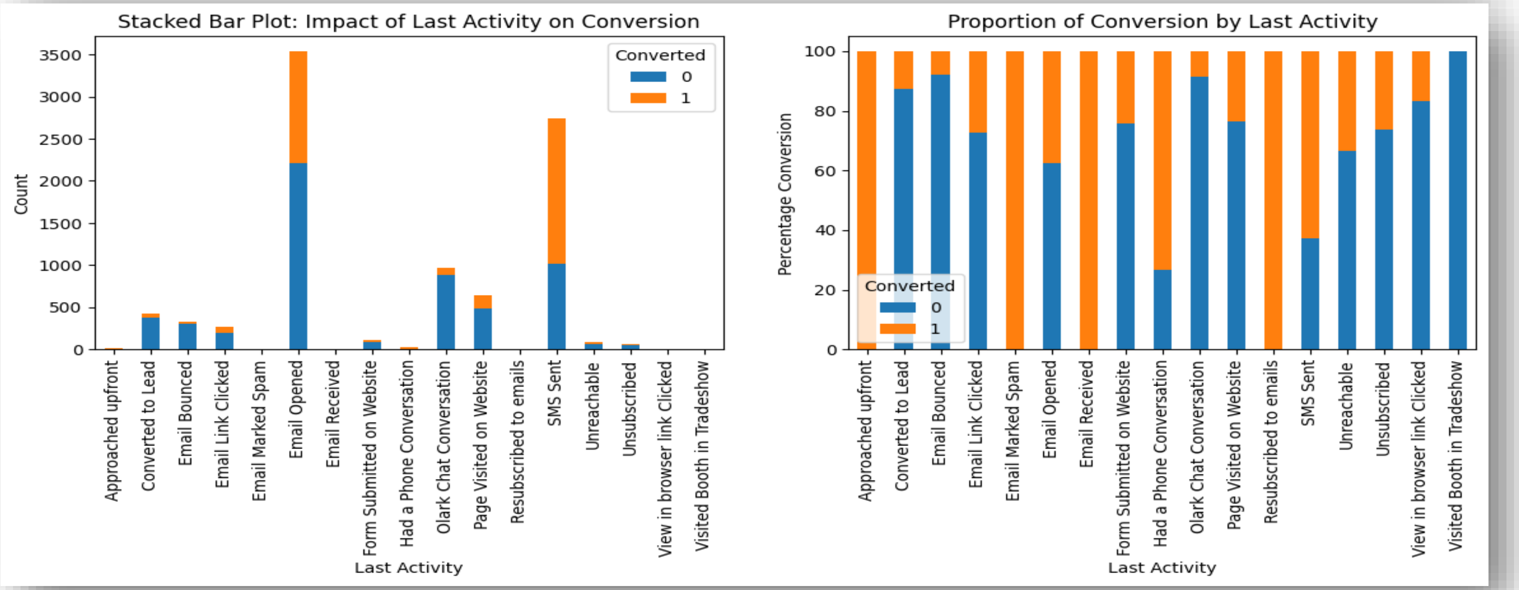
✓ Impact of Occupation on Conversion

- While Unemployed in the biggest pool in terms of leads conversion rate Working professionals is the best group to target.
- Housewives have 100% conversion rate but their numbers are low



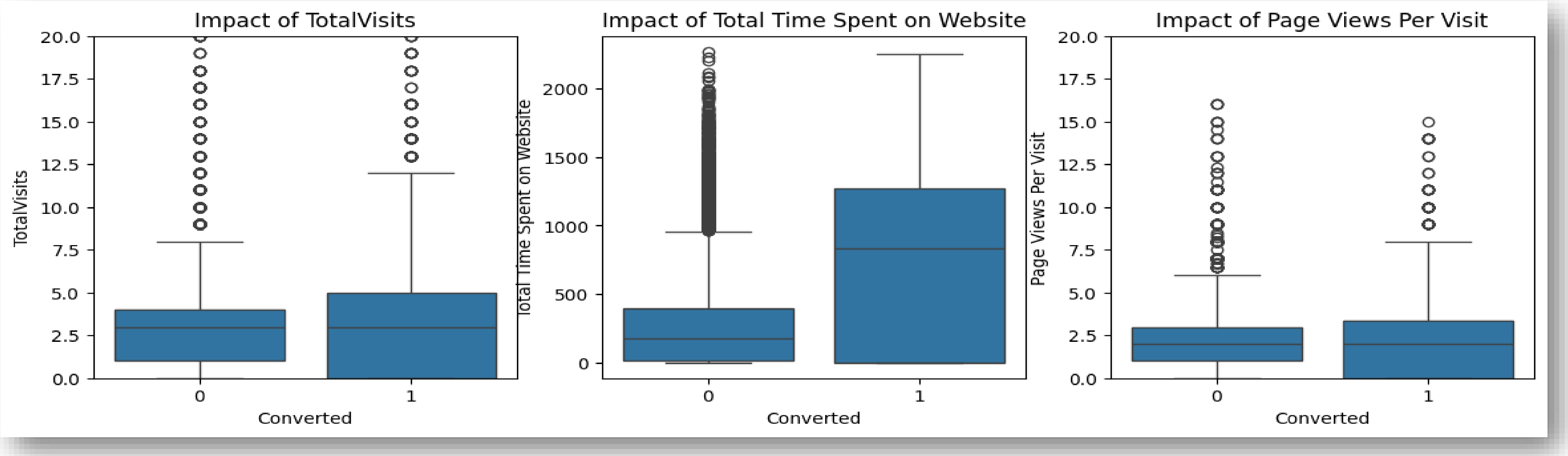
✓ Impact of Last Activity on Conversion

- Two major categories SMS Sent and Email Opened. SMS Conversation seems more promising than email in terms of percentage



Explanatory Data Analysis (3)

Numerical variables



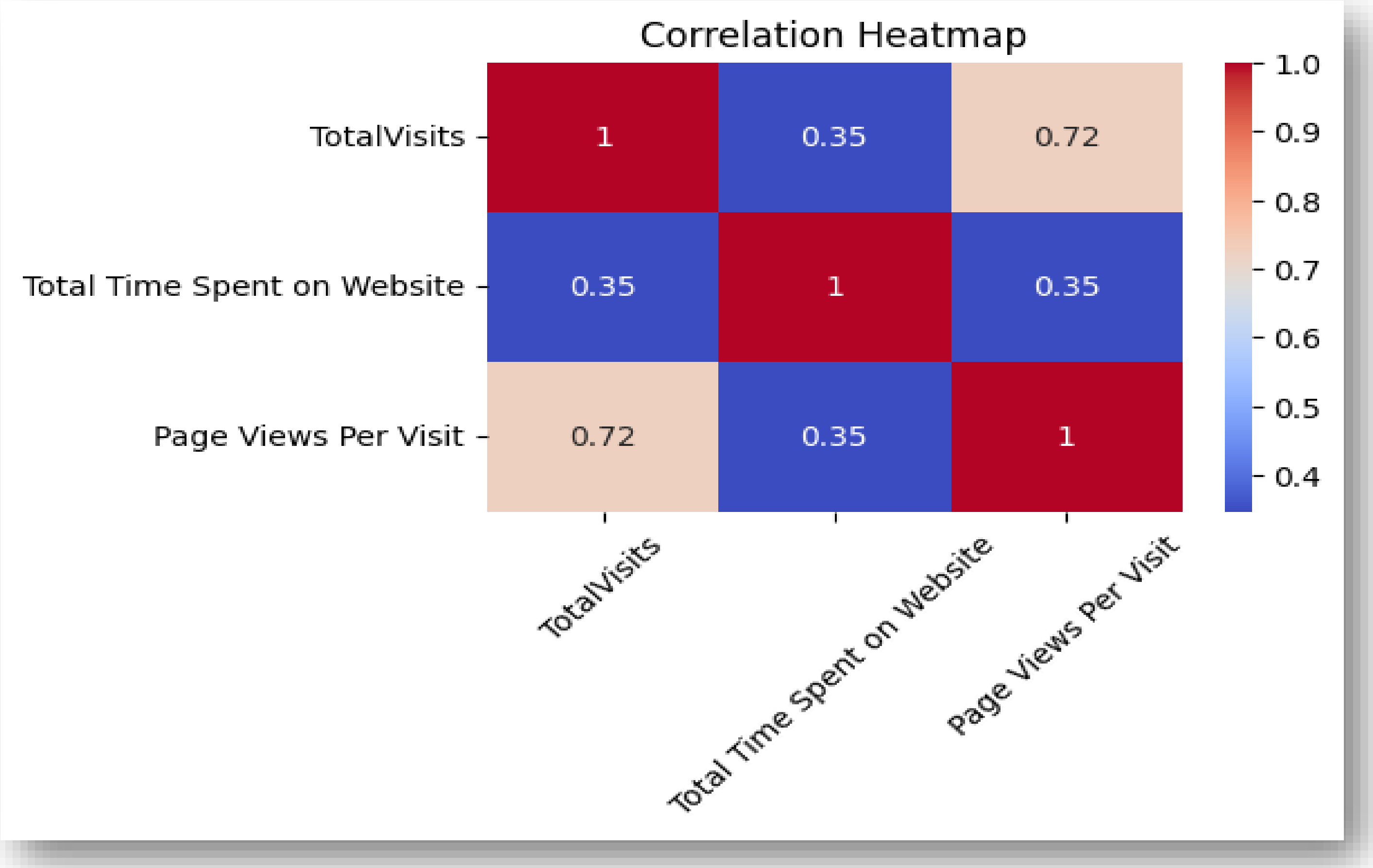
- Only time spent on website has impact on the conversion, Visits and page views median and quartiles are nearly same for conversion.
- People who spend more time are more interested.

❖ Insight - making the website more appealing may help

Explanatory Data Analysis (4)

Numerical variables

- Total Visits and Page Views have slightly higher correlation (0.72)
- Decided to retain the variables as the correlation is not that high and VIF will used post the models to remove any multi collinearity

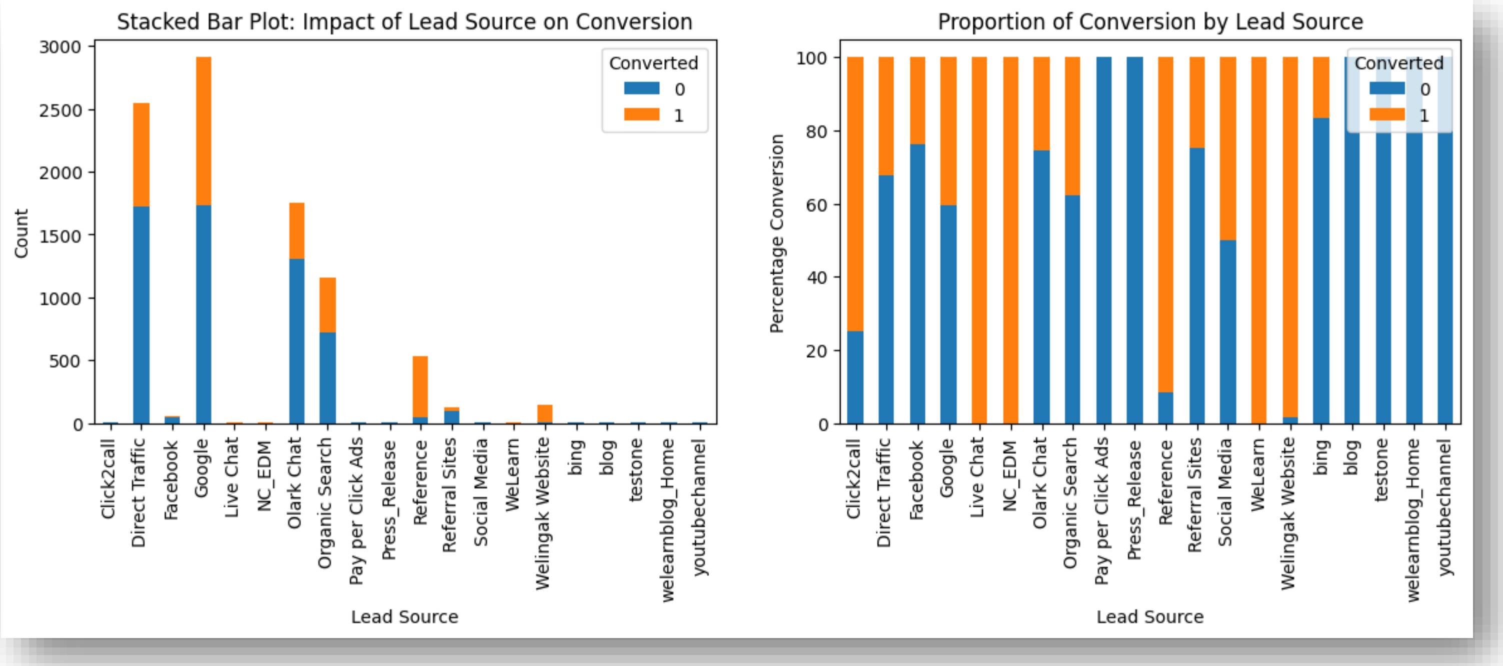


Explanatory Data Analysis (5)

Categorical variables

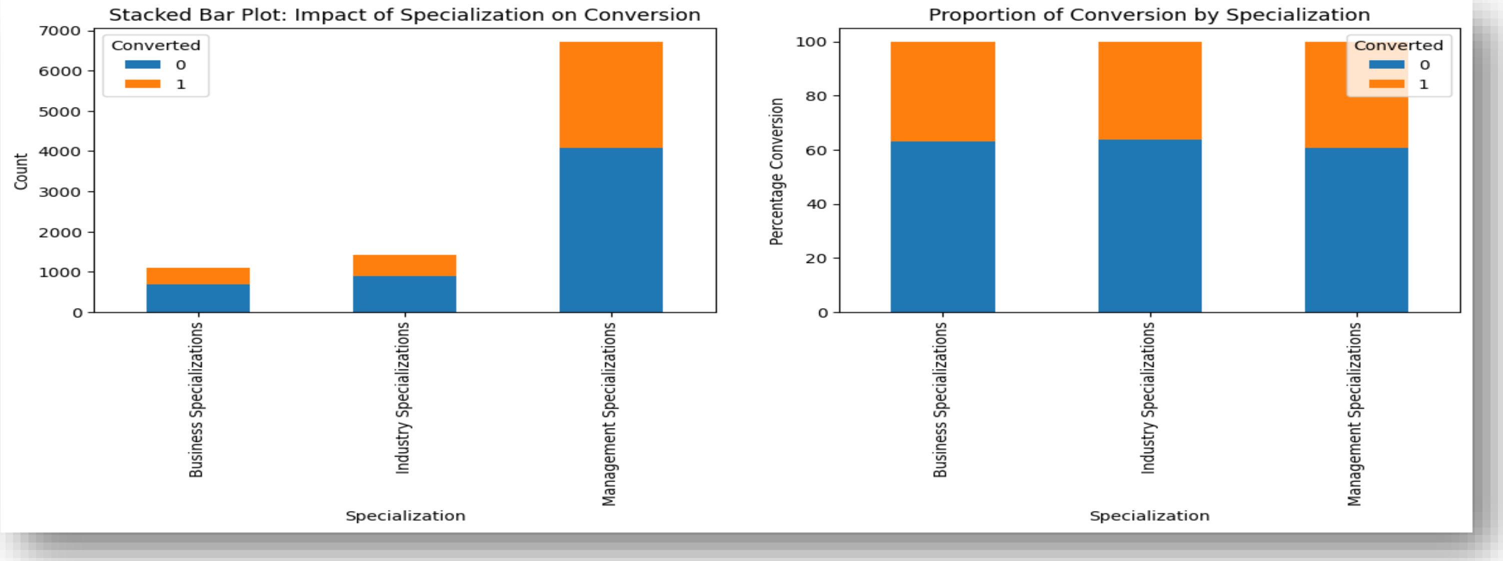
✓ Impact of Lead Source on Conversion

- While ‘Google’ and ‘Direct Traffic’ provide the maximum leads, ‘References’, ‘Welingak Website’ generate leads with decent number and best conversion rate
- ‘We learn’,’NC_EDM’,’Live Chat’ have 100% conversion, but their counts are lows



✓ Impact of Specialization on Conversion

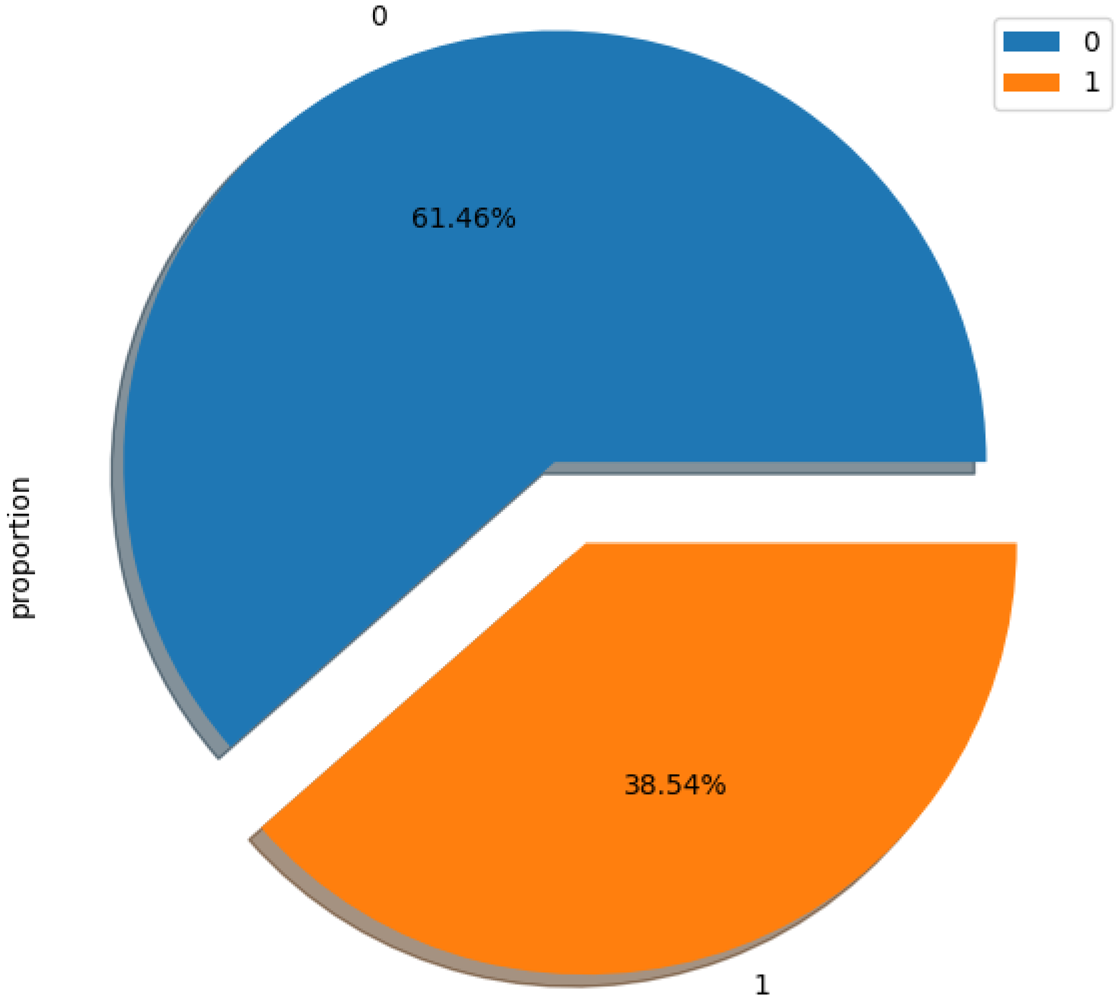
- Management Specialization gets the maximum leads but there is no difference in the percentage of conversion based on the type of Specialization



Explanatory Data Analysis (6)

Conversion Rate

- ✓ Checking the conversion rate in the Leads data provided
  - The conversion is on expected lines as stated in the problem statement
  - 39% Conversion rate means the data is balanced and not very skewed, Class imbalance is low



Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- As there were so many variables, RFE was used to select the top 15
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Identification of cut off/threshold
- Validation on test data set

Output of RFE

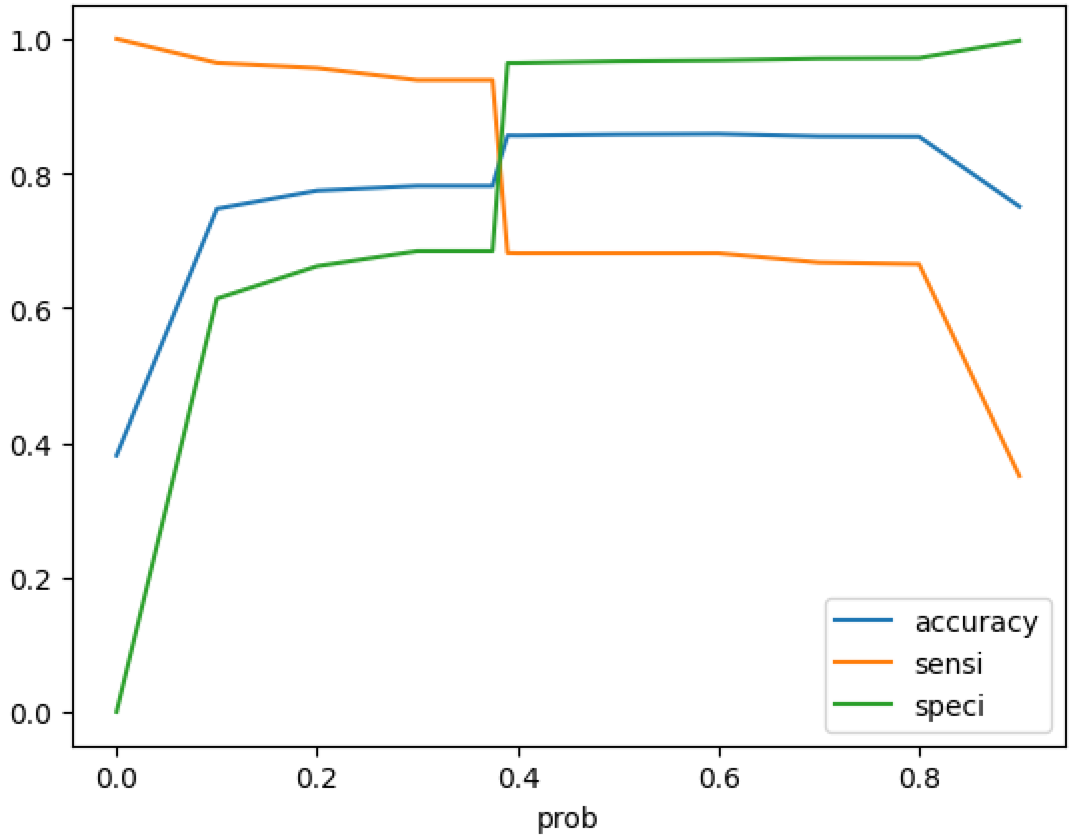
['Do Not Email', 'Lead Origin_Lead Add Form','Last Activity_Olark Chat Conversation', 'Occupation_Working Professional', 'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Interested in Next batch','Tags_Lateral student', 'Tags_Lost to EINS', 'Tags_Ringing','Tags_Will revert after reading the email', 'Tags_in touch with EINS’, 'Tags_switched off', 'Last Notable Activity_Had a Phone Conversation','Last Notable Activity_SMS Sent']

Cut off/ Threshold Identification

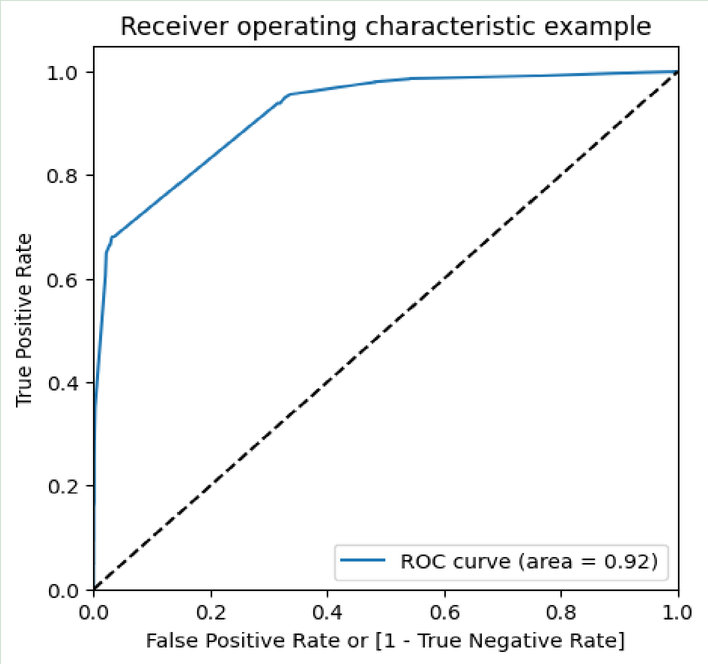
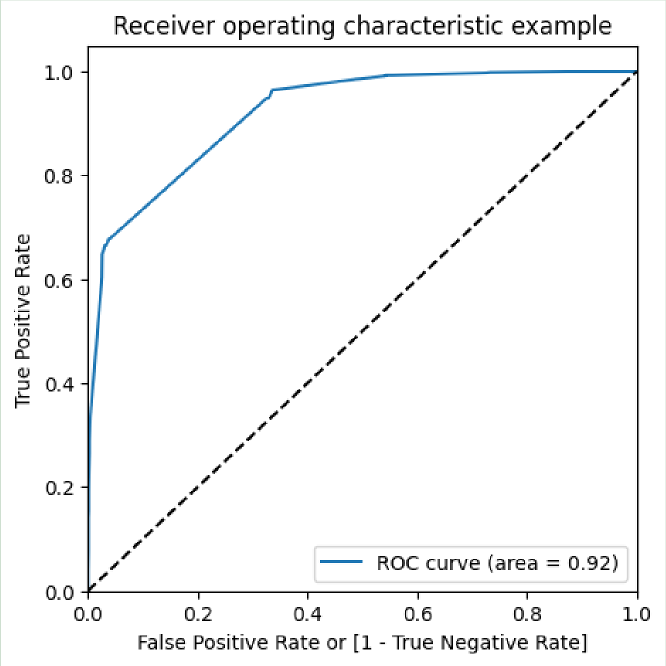
prob	accuracy	sensi	speci
0	0.381262	1	0
0.1	0.747372	0.964315	0.613693
0.2	0.774273	0.95661	0.661919
0.3	0.781385	0.938767	0.684408
0.34	0.781385	0.938767	0.684408
0.35	0.781385	0.938767	0.684408
0.36	0.781385	0.938767	0.684408
0.37	0.781385	0.938767	0.684408
0.375	0.781385	0.938767	0.684408
0.39	0.85637	0.681671	0.964018
0.4	0.856215	0.681265	0.964018
0.5	0.857916	0.681265	0.966767
0.6	0.858689	0.681265	0.968016
0.7	0.855133	0.667478	0.970765
0.8	0.854669	0.665045	0.971514
0.9	0.750773	0.35077	0.997251

Based on the metrices the threshold of **<u>0.375</u>** was selected

*Since its identification of leads, its better to identify more leads with few misses than miss a few leads so **<u>Sensitivity</u>** is key here*



Evaluate Train and Test sets on selected threshold/cut off – 0.375

Metrices	Train Set	Test Set
Accuracy	0.78	0.78
Sensitivity	0.93	0.94
Specificity	0.68	0.67
Precision	0.64	0.65
ROC Curve		

- ❖ The metrices have not varied much between Train and test sets
- ❖ Sensitivity is key here as cost of missing a lead is high. We can achieve 93% Sensitivity and 78% accuracy
- ❖ AUC of the curve is 0.92 which is good (large auc implies better model)
- ❖ **LEAD SCORE** – can be derived as = 100 * Conversion probability
- ❖ **LEAD SCORE of cut off 37.5**

Feature Importance for insights

Coefficients with positive impact on Conversion

Tags_Closed by Horizzon	8.823691
Tags_Lost to EINS	8.023987
Last Notable Activity_Had a Phone Conversation	3.541627
Tags_Will revert after reading the email	3.371749
Occupation_Working Professional	2.854458
Tags_Busy	2.790407
Tags_in touch with EINS	2.723909
Last Notable Activity_SMS Sent	2.508968
Lead Origin_Lead Add Form	2.415679

Coefficients with Negative impact on Conversion

Last Activity_Olark Chat Conversation	-3.881771
Tags_switched off	-1.507874
Do Not Email	-1.414326
Tags_Ringing	-1.342857

- Tags - Closed by Horizzon and Lost to EINS are major postivies so Customer tagged in this class should be followed
- Leads in touch – ‘Tags_Will revert after reading the email’ and ‘Last Notable Activity_Had a Phone Conversation’ have high positive coefficient
- ‘Working Professional’ is most preferred occupation
- ‘Last Activity_Olark Chat Conversation’ is most negative and should be avoided
- Leads who have lost connection – “Tags_switched off” / “Tags_Ringing“ don’t convert no point of pursuing them
- Leads who have provided negative feedback – “ Don’t Email” shouldn’t be chased

Conclusion

It was found that the variables that mattered the most in the potential buyers are

- ✓ Tags provide good information - Closed by Horizzon, Lost to EINS, Will revert after reading the email have high positive impact
- ✓ Leads who are in touch - Last Notable Activity_Had a Phone Conversation, Last Notable Activity_SMS Sent - convert more
- ✓ Best the lead source in terms of counts – Google, Direct TrafficWelingkar website
- ✓ Best lead source in terms of conversion rate - ‘References’, ‘Welingak Website’
- ✓ Working Professionals have highest conversion rate
- ○ Last Activity_Olark Chat Conversation should be avoided
- ○ Leads who don’t respond – switched off, ringing, Tags- Don’t Email should be avoided have most negative impact