Lead Scoring Case Study

DSC 55 Batch

Submission on 16-10-2023

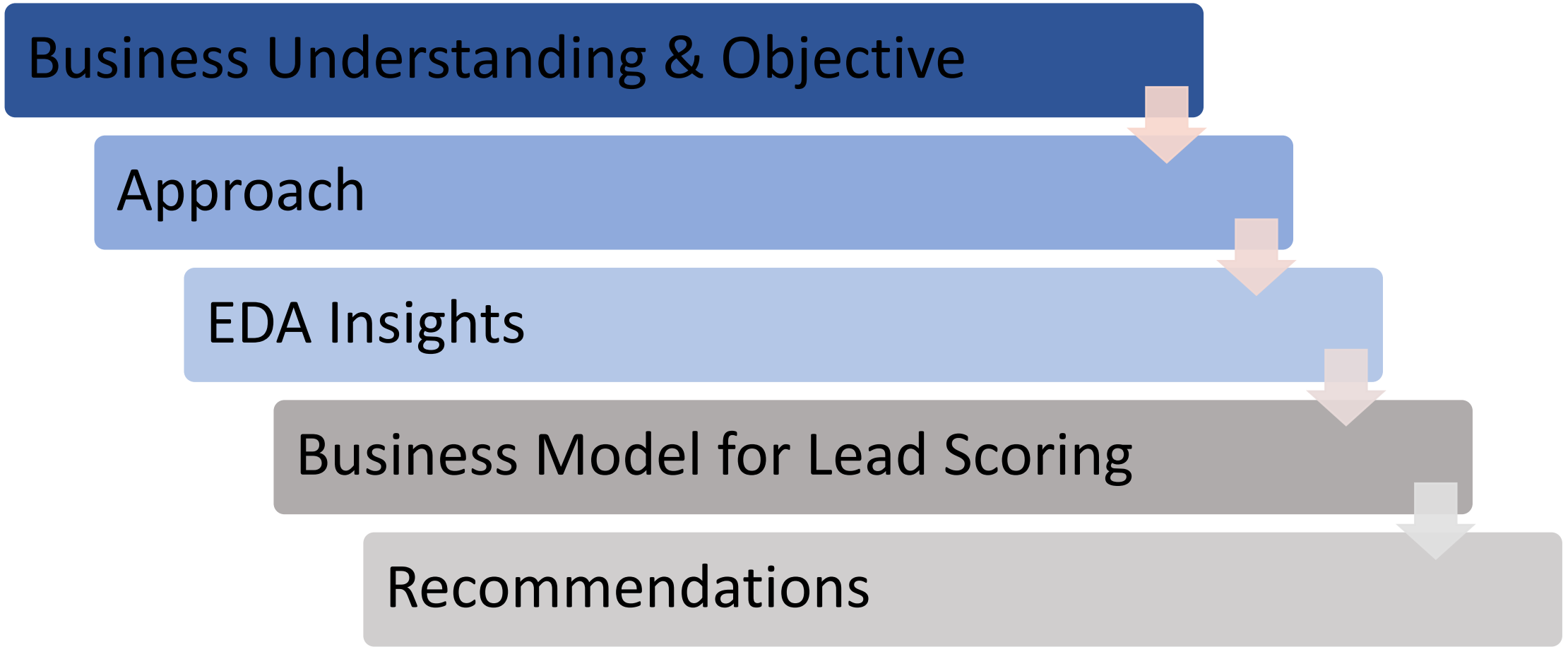
by

Prem Prakash

Priya G Rao

Pritha Suresh

Schema


Business Understanding & Objective
Approach
EDA Insights
Business Model for Lead Scoring
Recommendations

Business Understanding and Objective

X Education, an education company selling online courses wants to improve its lead conversion rate.

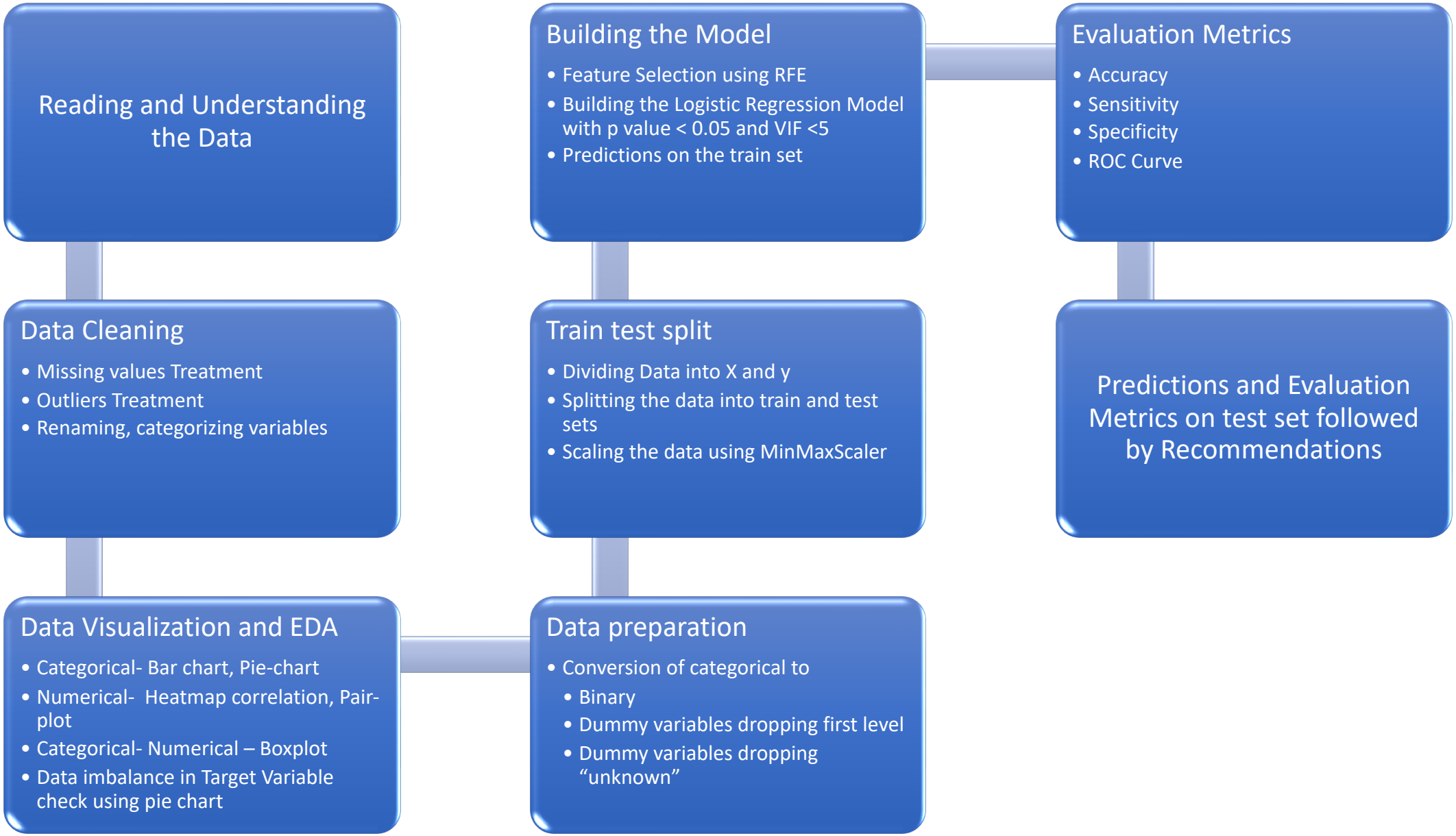
Present Scenario
Lead Conversion rate : 30%
Lead Sources:

- Social Media
- Marketing
- Past Referrals

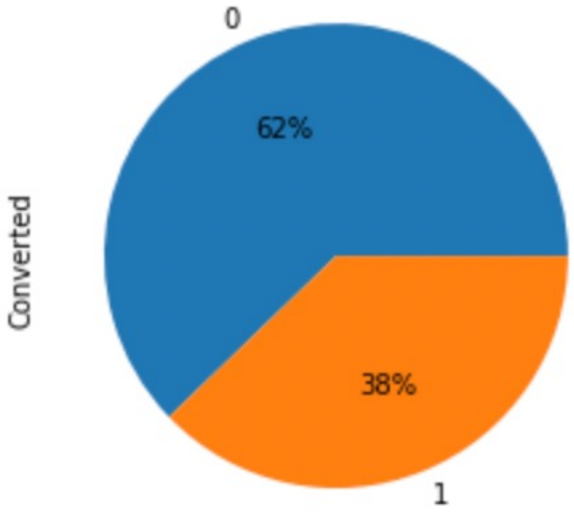
Objective:
The Company wants to identify its most potential leads known as "hot leads" thereby increase its conversion rate by focusing on these leads. The CEO has set a target lead conversion rate of 80%.

Requirement:
The task is to build a Logistic Regression Model that will assign a lead score between 0 to 100 to each of the 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.

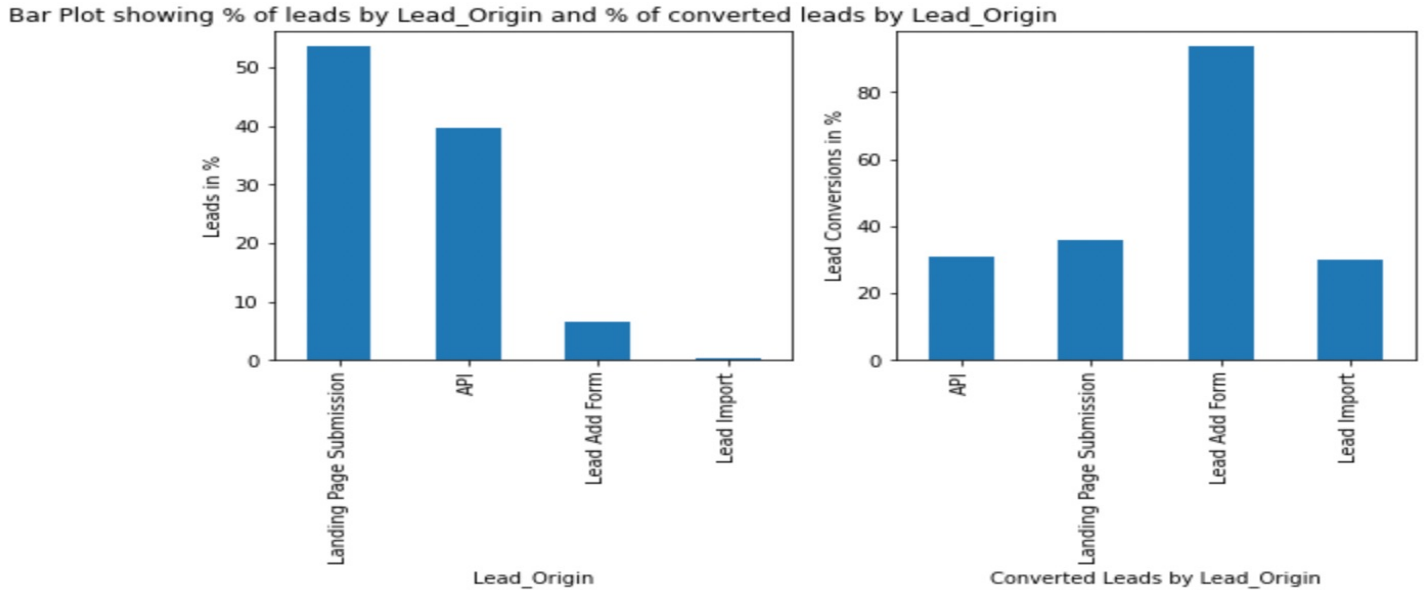
Approach

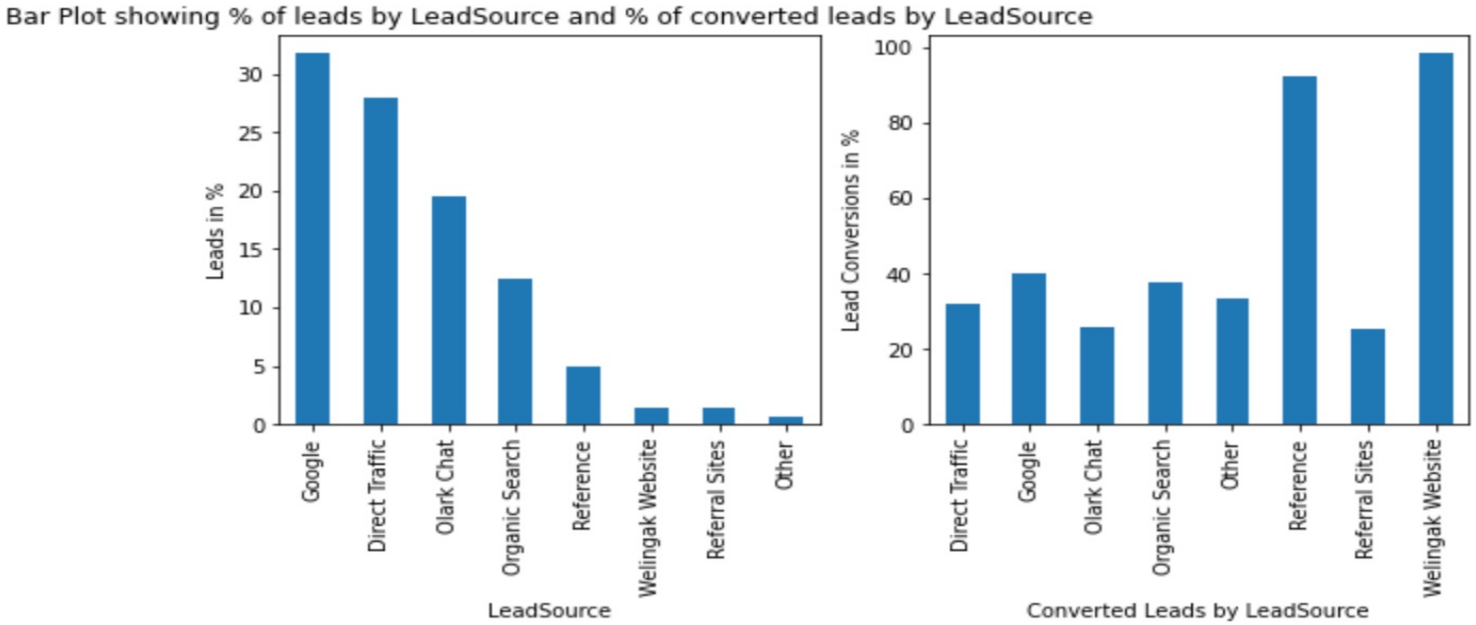

Reading and Understanding the Data
Data Cleaning
• Missing values Treatment
• Outliers Treatment
• Renaming, categorizing variables
Data Visualization and EDA
• Categorical- Bar chart, Pie-chart
• Numerical- Heatmap correlation, Pair-plot
• Categorical- Numerical – Boxplot
• Data imbalance in Target Variable check using pie chart
Data preparation
• Conversion of categorical to
• Binary
• Dummy variables dropping first level
• Dummy variables dropping "unknown"
Train test split
• Dividing Data into X and y
• Splitting the data into train and test sets
• Scaling the data using MinMaxScaler
Building the Model
• Feature Selection using RFE
• Building the Logistic Regression Model with p value < 0.05 and VIF <5
• Predictions on the train set
Evaluation Metrics
• Accuracy
• Sensitivity
• Specificity
• ROC Curve
Predictions and Evaluation Metrics on test set followed by Recommendations

EDA Insights

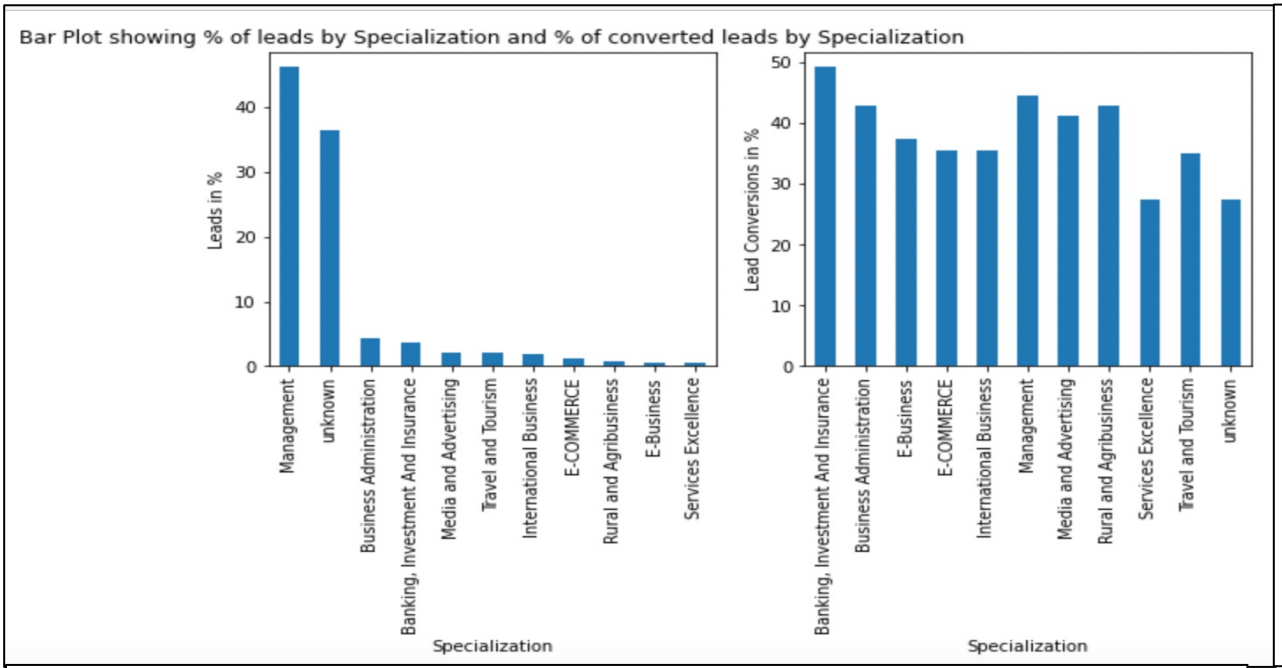


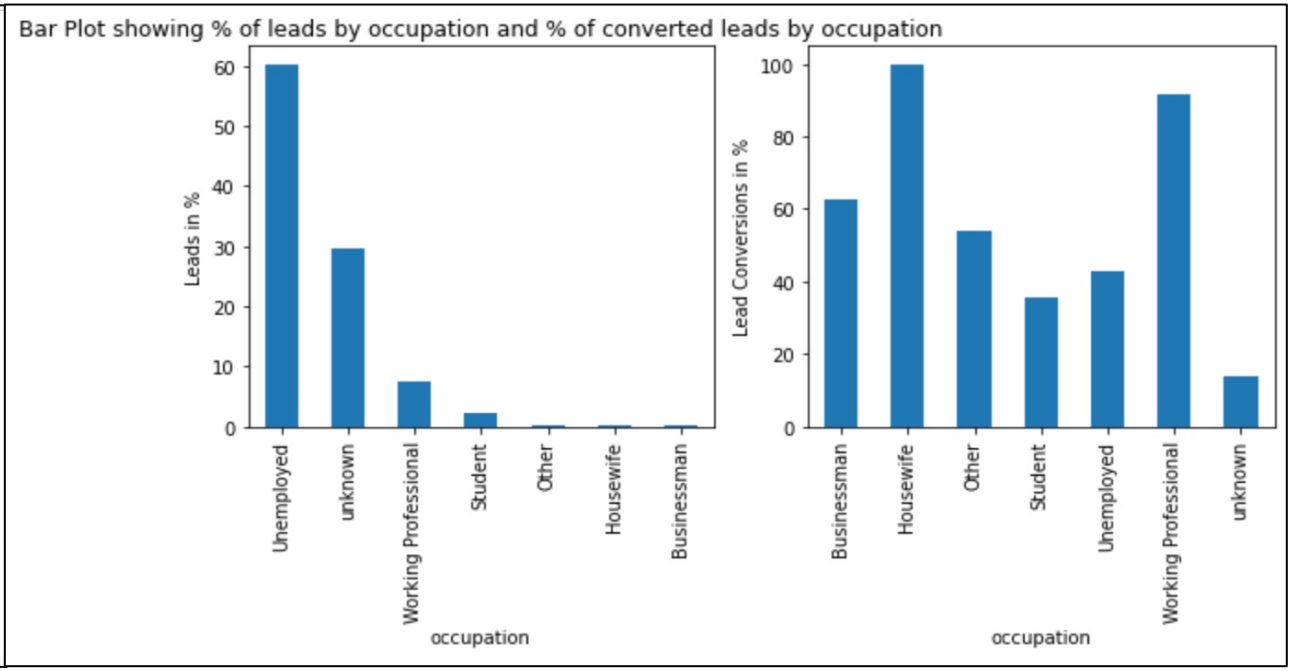
- The Target Variable ”Converted” has a good mix of converted and not converted leads.
- Conversion rate in the data used for building the model is 38%.
- Though most leads are from Landing Page Submission, the lead conversion rate is higher in Lead Add Form
- Welingak Website and Reference have very high conversion rates of more than 90%.

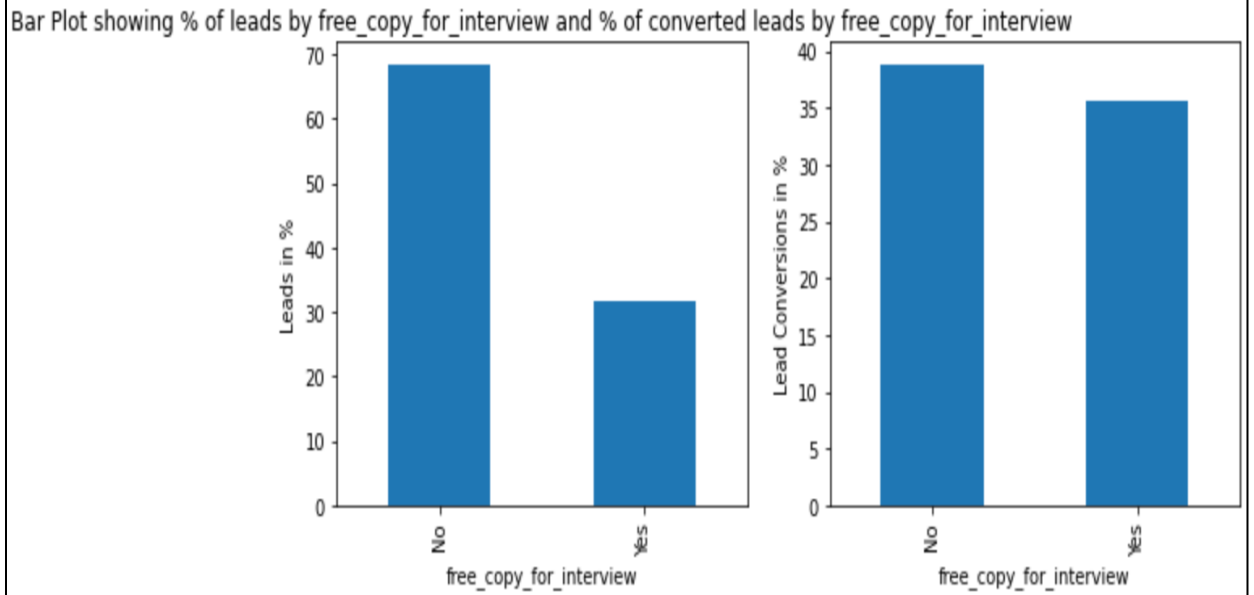




EDA Insights

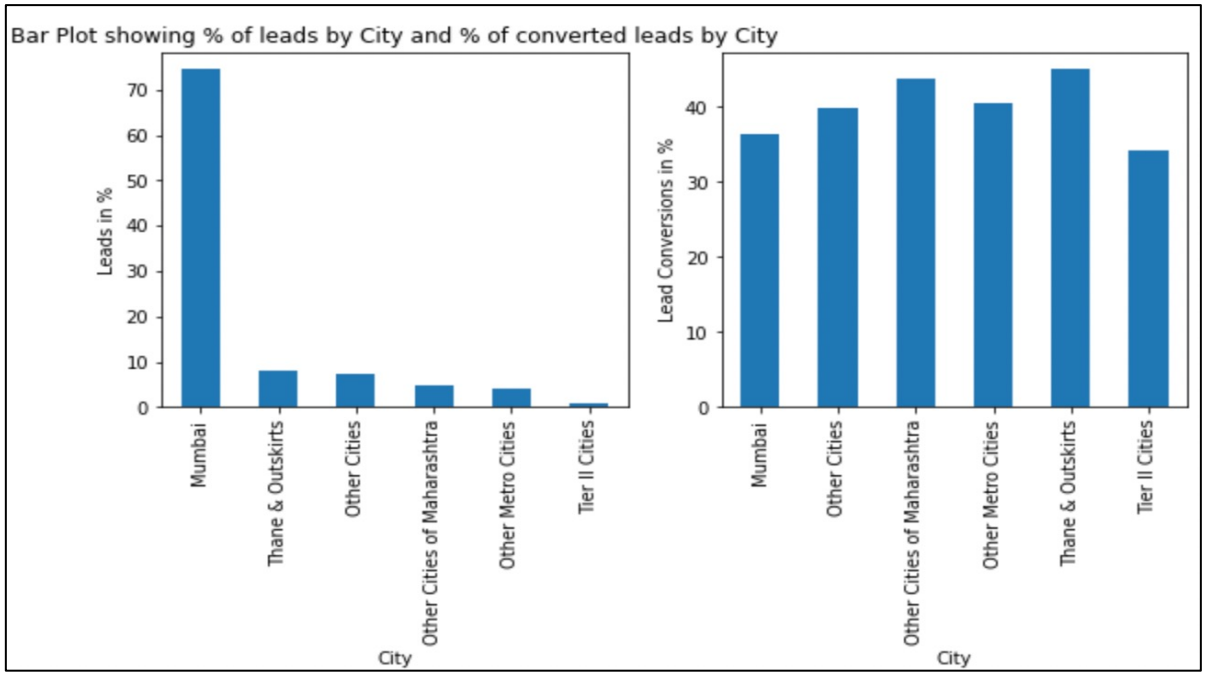


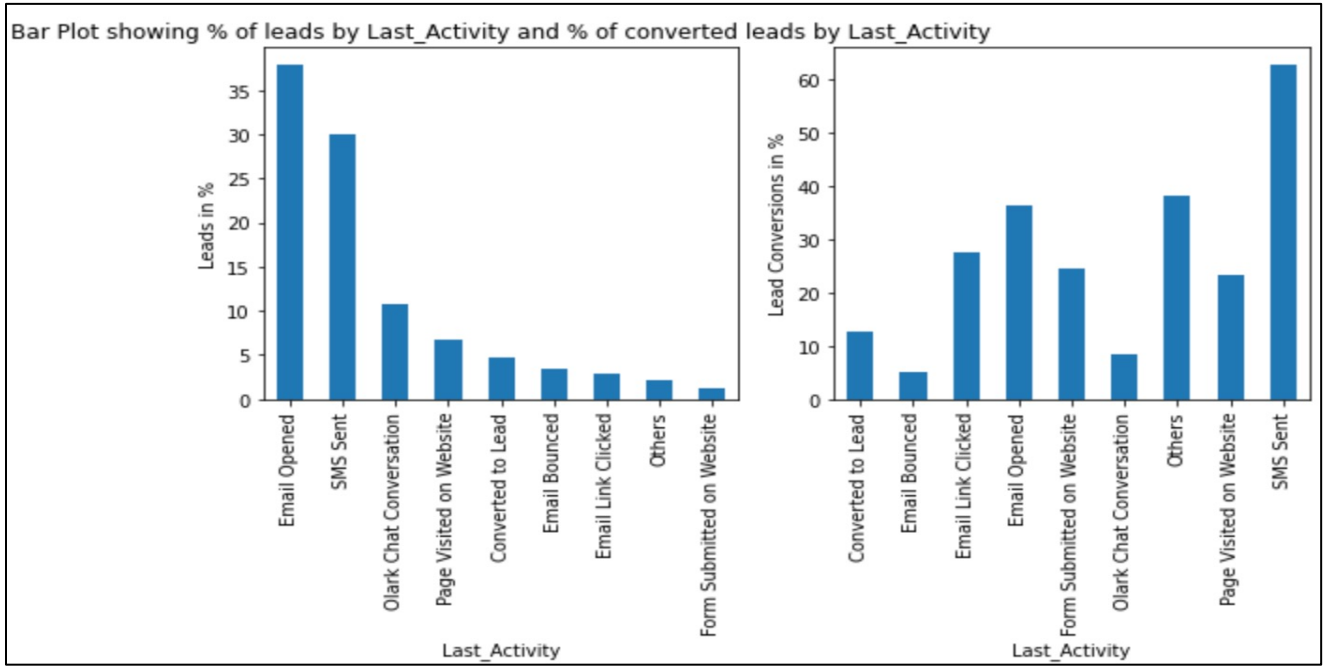


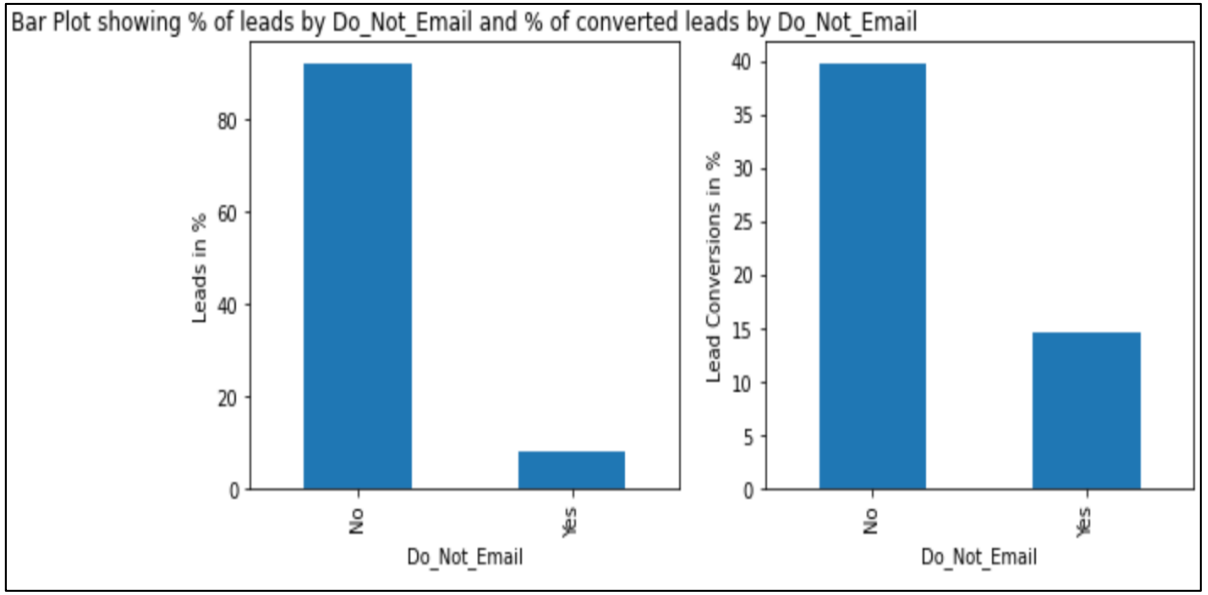


- Majority of the leads are from persons of Management streams followed by persons whose Specialization is not known.
- Banking Investment and Insurance has higher conversion rate
- Though leads comprising of Housewife and Working Professionals are less than 10%, they have very high very conversions of 100% and 94% respectively
- Conversion rate is more than 35% irrespective of whether the leads opted or not for the “free copy for interview”.

EDA Insights

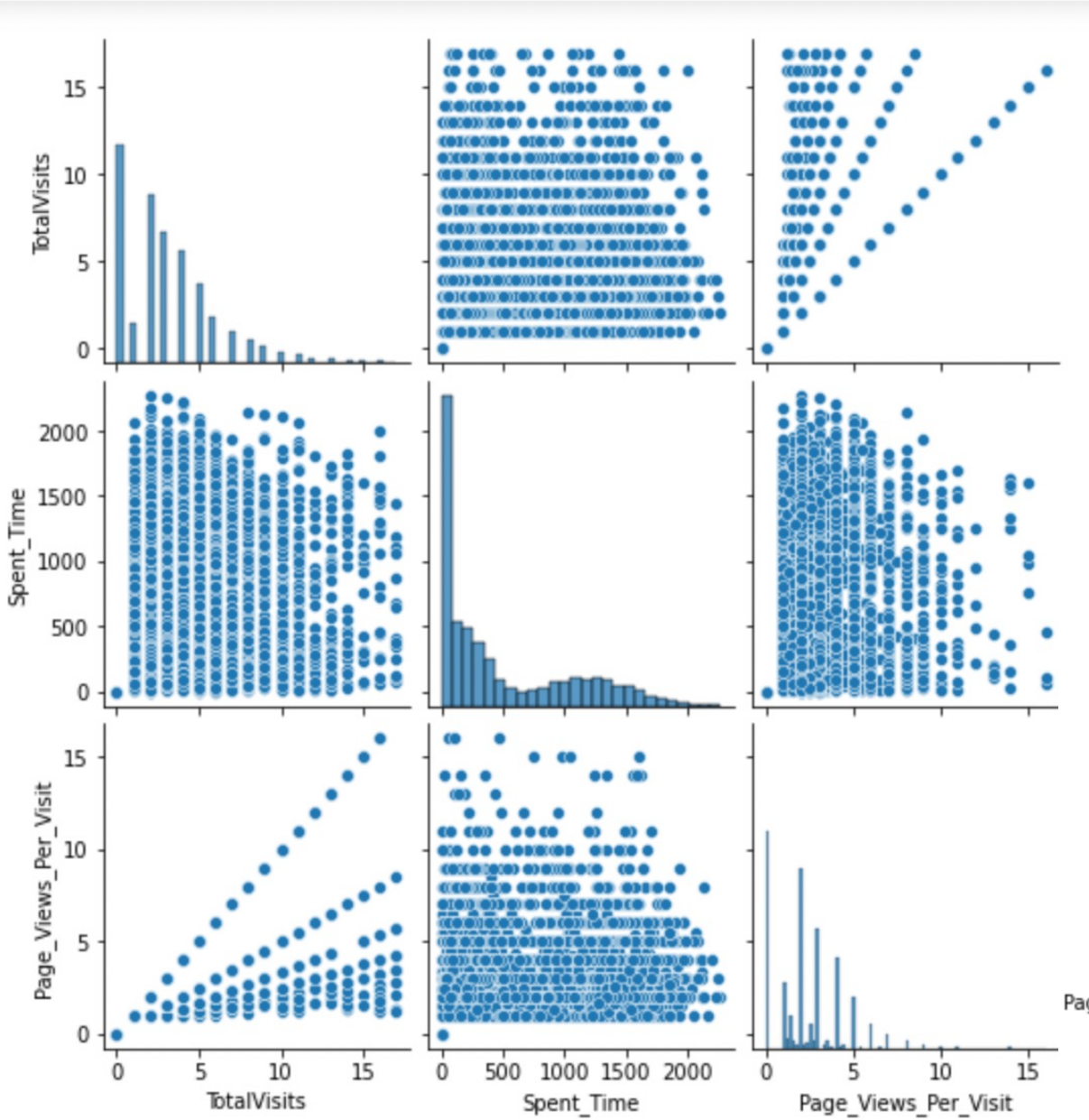




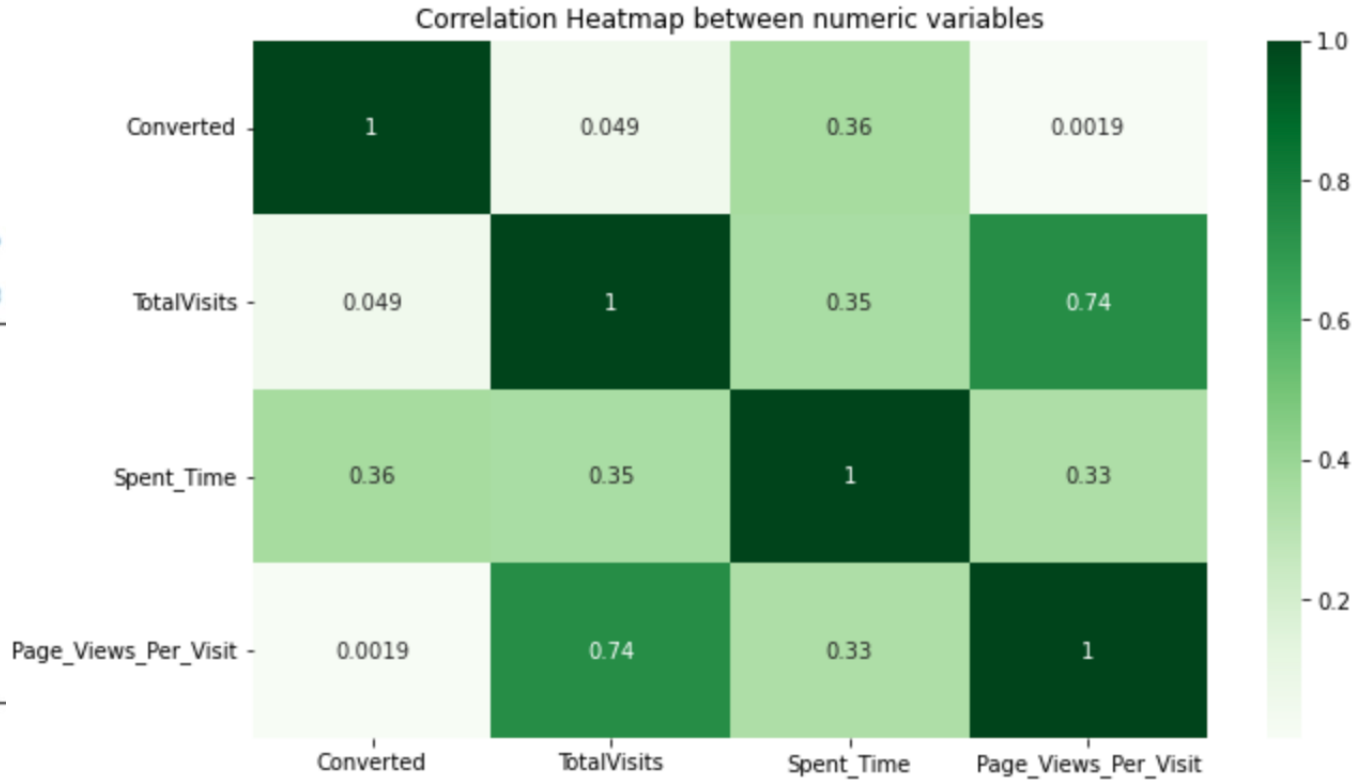


- Conversion rates are in the range of 35-45% across Cities.
- Conversions are high where the Last_Activity is SMS sent.
- Leads who had requested for Emails have opted for the courses in almost 40% of the cases.

EDA Insights

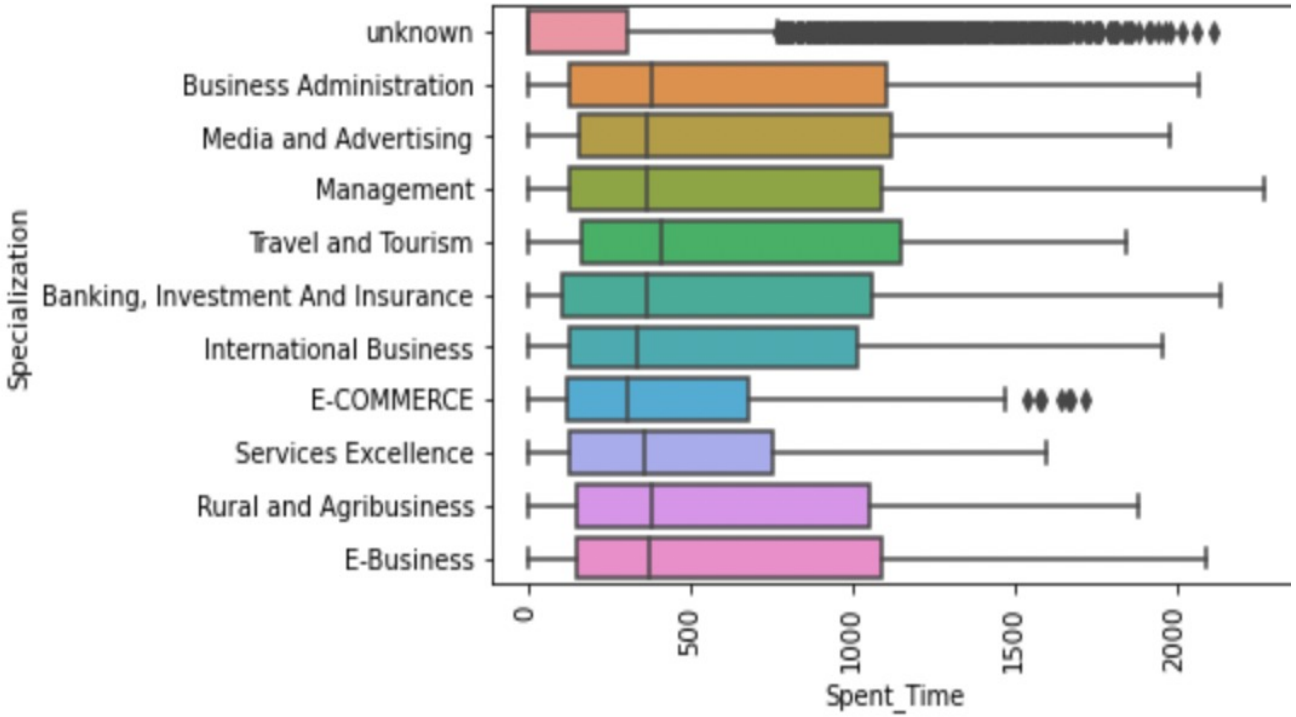


There exists a positive correlation between TotalVisits and Page_Views_Per_Visit.

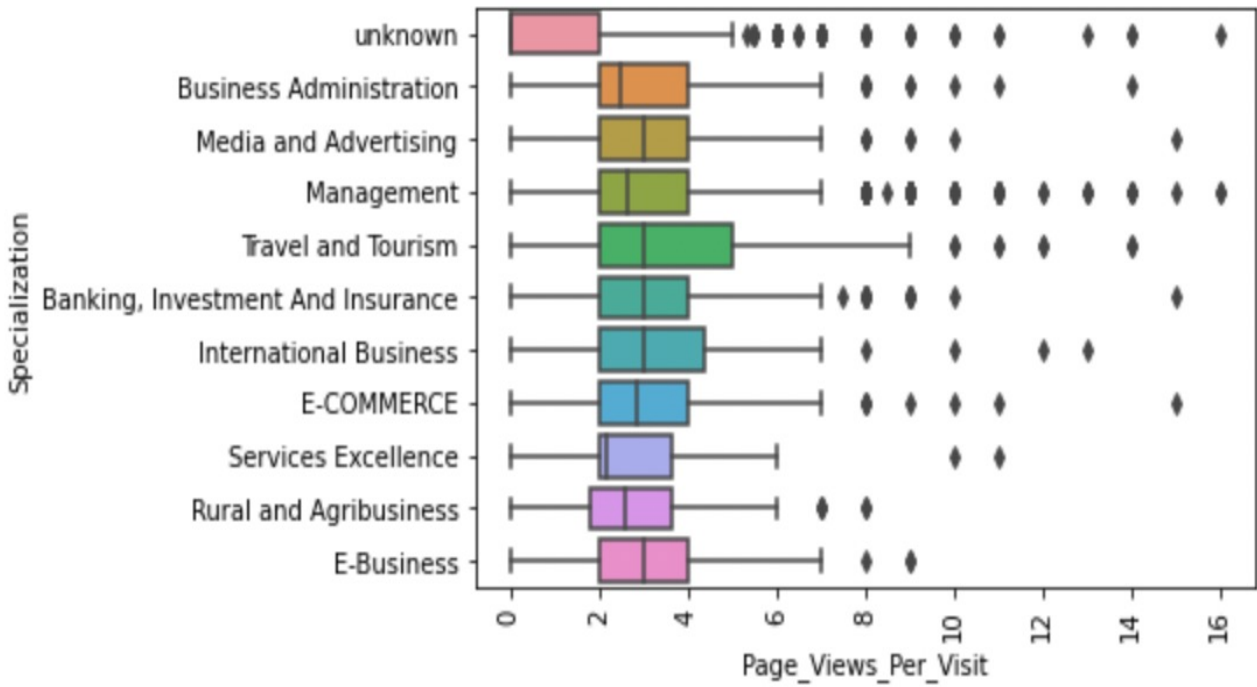


EDA Insights

Boxplot Spent_Time Vs Specialization



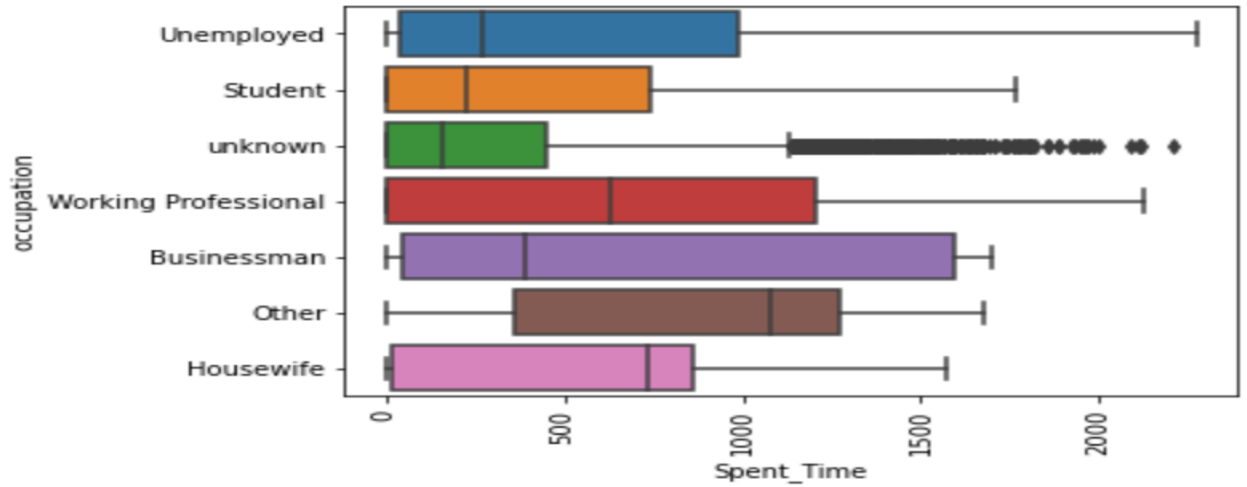
Boxplot Page_Views_Per_Visit Vs Specialization



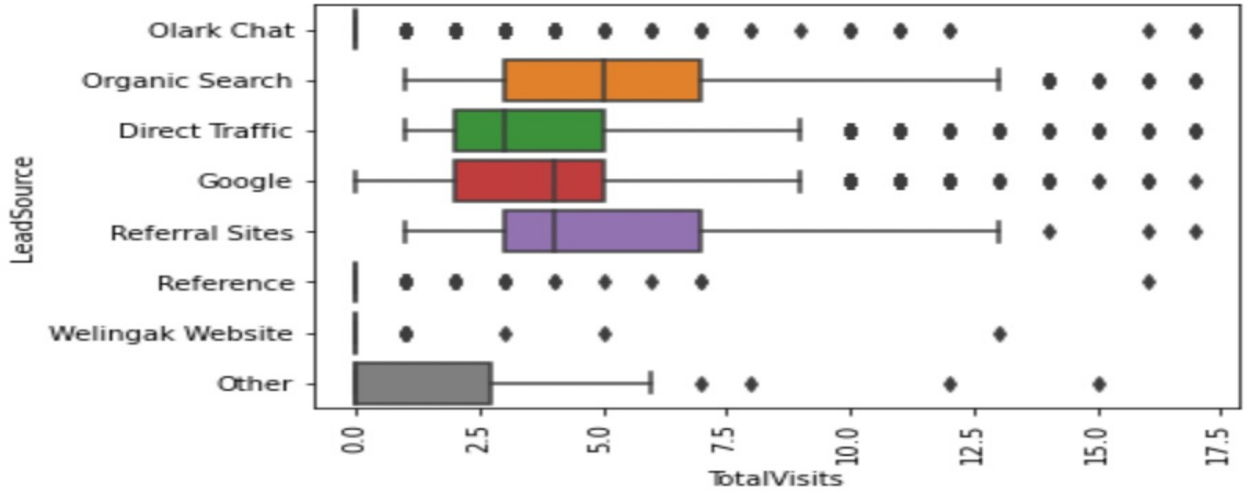
- Though Management stream leads spend more time on the website , the page views per visit are higher for leads coming from Travel and Tourism Specialization

EDA Insights

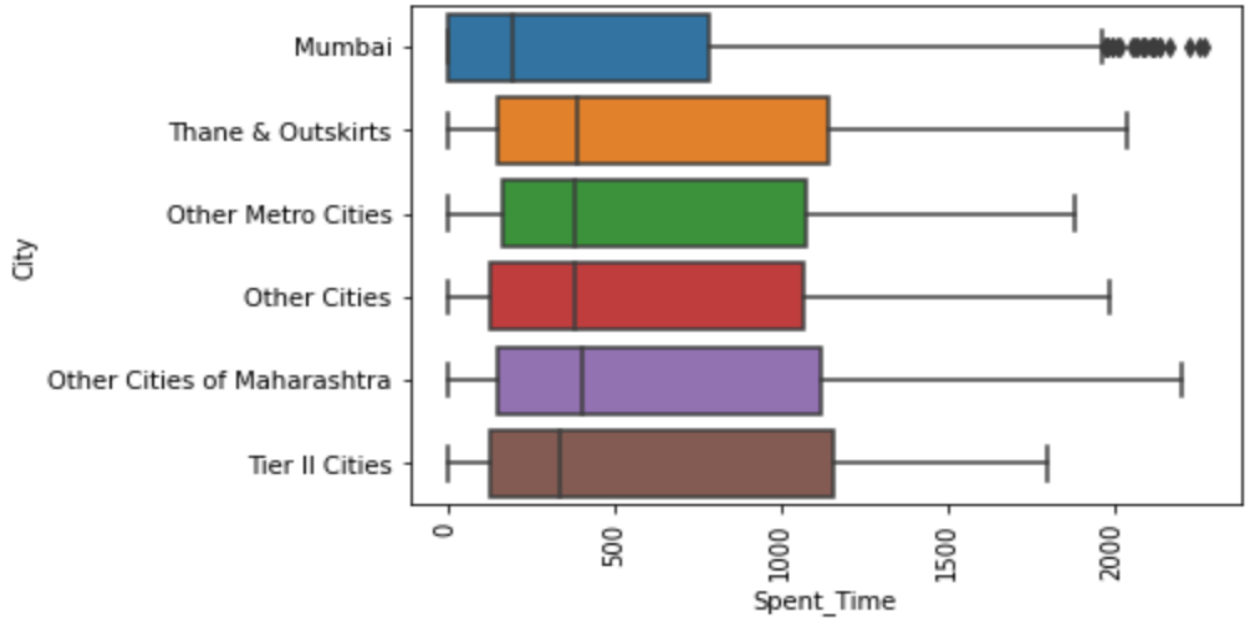
Boxplot Spent_Time Vs occupation



Boxplot TotalVisits Vs LeadSource



Boxplot Spent_Time Vs City



- Unemployed Persons spend more time checking information on the website
- Total Visits are more in Organic Search and through Referral Sites
- Persons in Other Cities of Maharashtra spend more time checking information on the website.
- There are few leads in the city of Mumbai who spend more time on the website.

Business Model Parameters

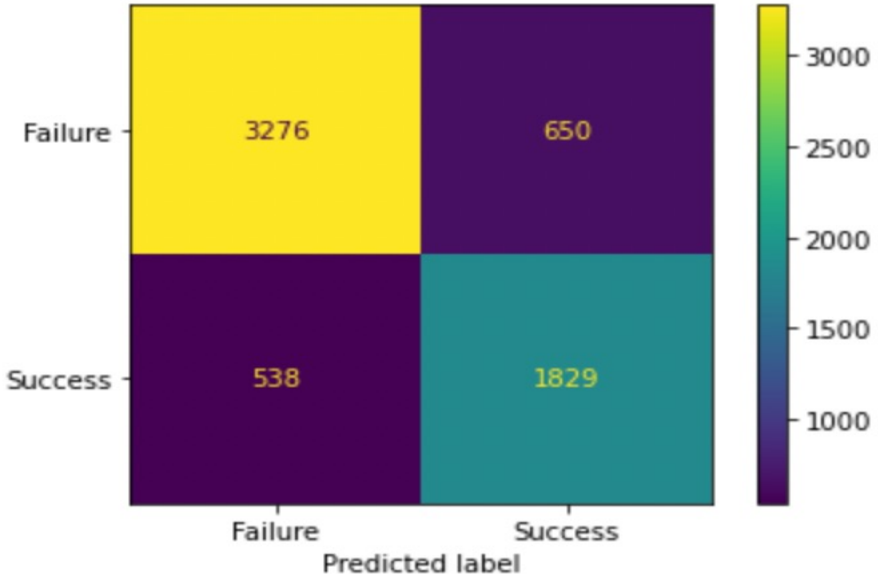
	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.5762	0.127	-28.138	0.000	-3.825	-3.327
Do_Not_Email	-1.5949	0.174	-9.183	0.000	-1.935	-1.254
TotalVisits	1.6304	0.286	5.698	0.000	1.070	2.191
Spent_Time	4.6622	0.169	27.519	0.000	4.330	4.994
Page_Views_Per_Visit	-1.3103	0.414	-3.161	0.002	-2.123	-0.498
Lead_Origin_Lead Add Form	4.2075	0.265	15.904	0.000	3.689	4.726
LeadSource_Olark Chat	1.5876	0.126	12.637	0.000	1.341	1.834
LeadSource_Welingak Website	2.2328	1.042	2.143	0.032	0.191	4.275
Last_Activity_Olark Chat Conversation	-1.2986	0.167	-7.779	0.000	-1.626	-0.971
Last_Activity_Others	0.9208	0.222	4.154	0.000	0.486	1.355
Last_Activity_SMS Sent	1.3676	0.076	17.889	0.000	1.218	1.517
occupation_Student	1.2624	0.235	5.362	0.000	0.801	1.724
occupation_Unemployed	1.1845	0.087	13.578	0.000	1.013	1.355
occupation_Working Professional	3.7261	0.202	18.444	0.000	3.330	4.122
Specialization_Banking, Investment And Insurance	0.4817	0.177	2.723	0.006	0.135	0.828

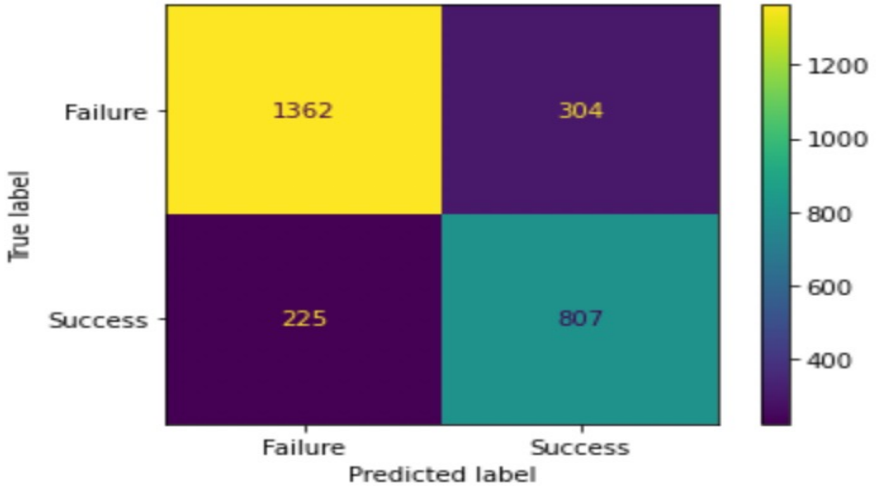
Features	VIF
const	9.621769
Do_Not_Email	1.037757
TotalVisits	2.327437
Spent_Time	1.311371
Page_Views_Per_Visit	2.574292
Lead_Origin_Lead Add Form	1.641182
LeadSource_Olark Chat	1.938879
LeadSource_Welingak Website	1.288301
Last_Activity_Olark Chat Conversation	1.276226
Last_Activity_Others	1.037105
Last_Activity_SMS Sent	1.121970
occupation_Student	1.054539
occupation_Unemployed	1.304182
occupation_Working Professional	1.297922
Specialization_Banking, Investment And Insurance	1.011985

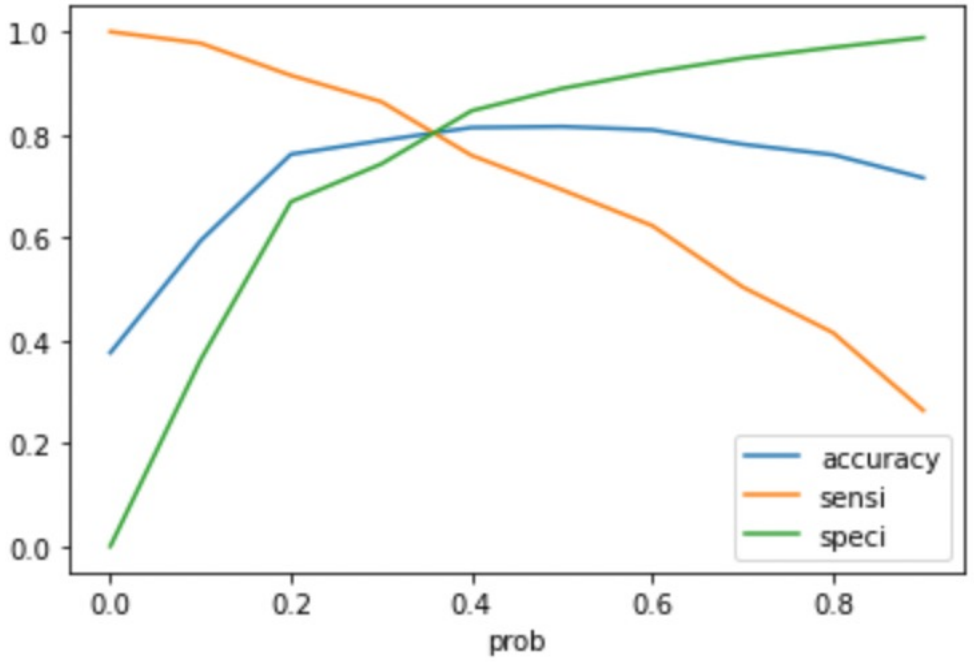
- The model has 14 parameters whose p value is less than 0.05 and VIF is less than 5
- The top 3 variables are Spent_Time, Lead_Origin_Lead Add Form and occupation_Working Professional
- Variables with meaningful insights in EDA are part of the final model

Business Model Evaluation Metrics

Confusion Matrix of train and test set







prob	accuracy	sensi	speci
0.0	0.376132	1.000000	0.000000
0.1	0.593834	0.977609	0.362455
0.2	0.761958	0.915082	0.669638
0.3	0.788495	0.863963	0.742995
0.4	0.813761	0.759611	0.846409
0.5	0.815668	0.692860	0.889710
0.6	0.809312	0.623152	0.921549
0.7	0.781344	0.504014	0.948548
0.8	0.761004	0.415294	0.969435
0.9	0.716034	0.264047	0.988538

- Since the target is to increase conversion rate, sensitivity is an important metric to evaluate
- The predicted class in final model is based on the optimal cut-off point of 0.38
- Train set has an accuracy, sensitivity and specificity of 81.12%,77.27% and 83.44% respectively
- Test set has an accuracy, sensitivity and specificity of 80.39%, 78.19% and 81.75% respectively.

Recommendations

- As a thumb rule ,X Education should focus on leads whose probability of conversions is more than the probability threshold of 0.40
  - Help in Cost savings to the company- less calls , more conversion
  - Increase motivation of the Sales force as the conversion rate would increase.
- Strategy when 10 sales interns are available:
  - Change the threshold limit for the hot leads to 0.05 to ensure that maximum calls are made .This would address leads which can actually be converted but which are predicted as “cannot convert”
  - Segment the Sales team into teams that will handle leads where probability of conversion is more than 0.40 and teams that will handle leads where probability is between 0.05-0.4. Communication can be tailored accordingly.
- Strategy when sales target met in advance
  - Focus only on leads where there is more than 0.5 probability of conversion so that calls are productive.
  - Put helpful content on the website so that leads can get the required information
- Others
  - Customize communication package for different segments of leads
    - For example- Working professional may not like being called during working hours.
    - Placement options for Unemployed and students as part of the course
  - More ads on Welingak Website
  - Arrange webinars, communication packages