X Education - Lead Scoring Case Study

Identify the most potential leads, also known as 'Hot Leads' to improve the lead conversion rate.

By-
Anannya Nair
Gaurang Sharma

About the company

- X Education, an Education company 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.

Business model

- 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%.

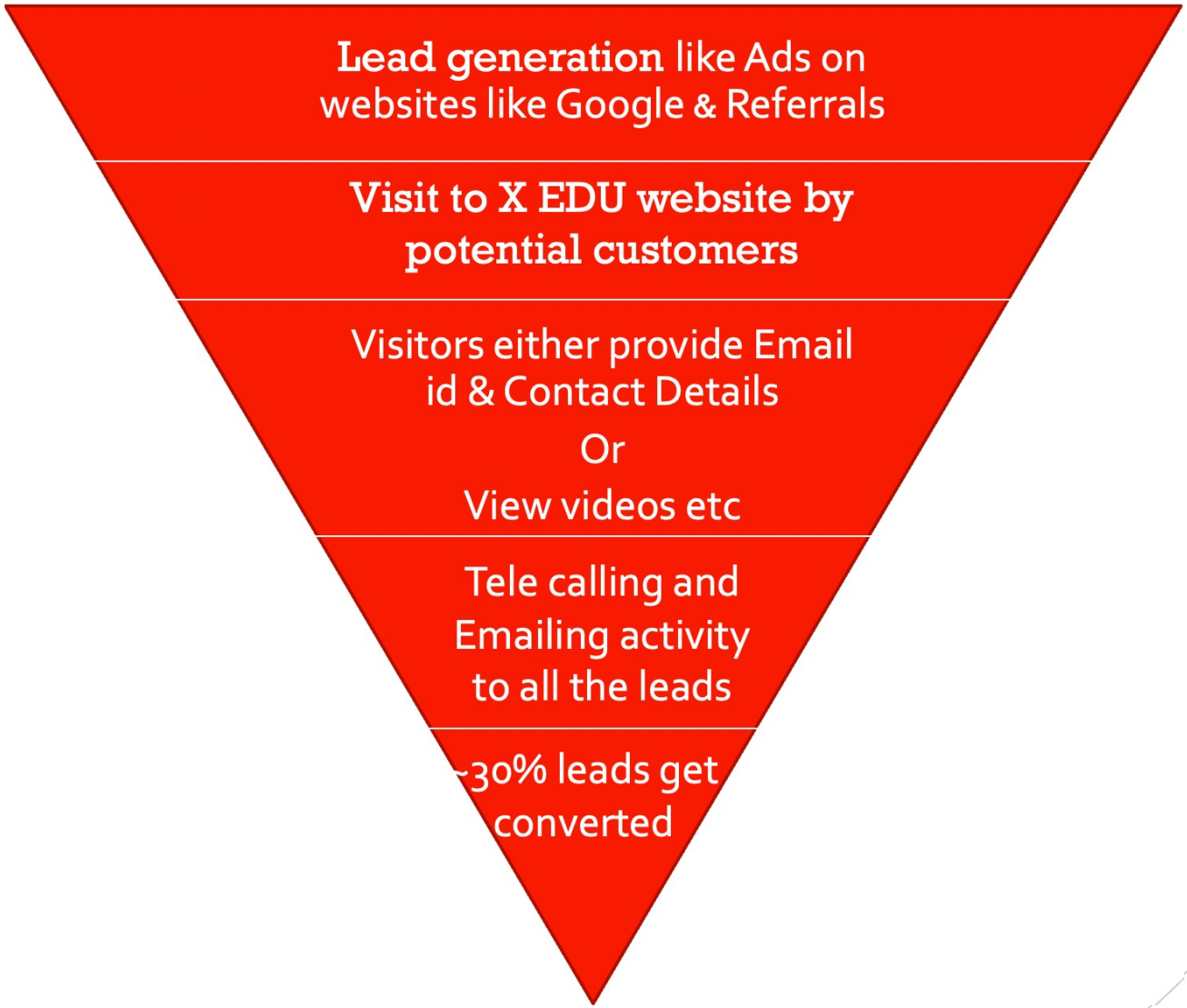
Problem Statement

- X Education gets a lot of leads, its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- On successfully identifying these sets of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads.

Our job

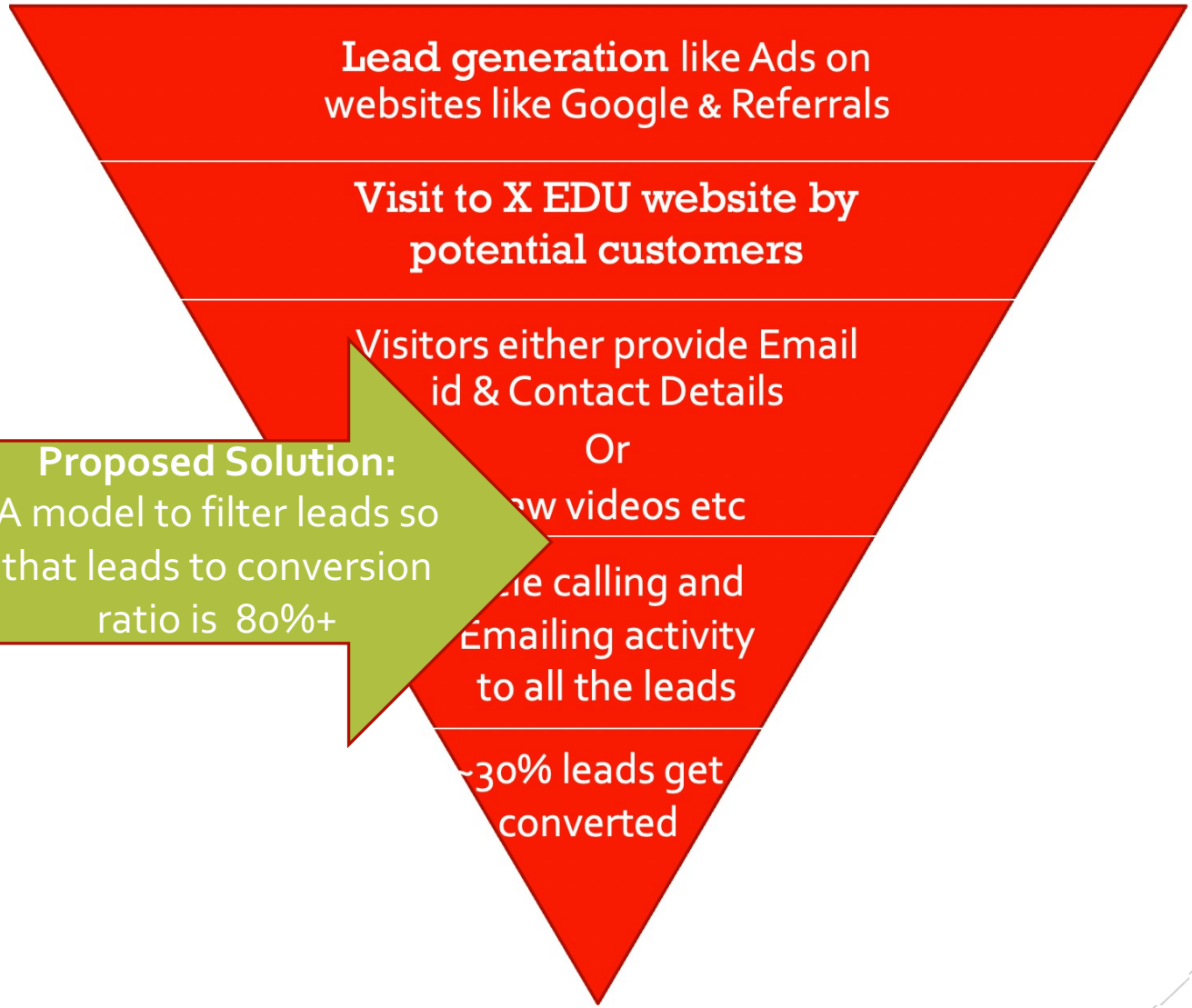
- X Education needs our help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We will build them a model wherein a lead score will be assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. Easy identification of potential customers.
- We have been asked to get their target lead conversion rate to be around 80%.

Current process



Fitting our model

Proposed Solution:
A model to filter leads so that leads to conversion ratio is 80%+



Proposed Solution

ion of Hot Leads &
unicating with Hot

Leads Clustering

We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

Focus Communication

Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

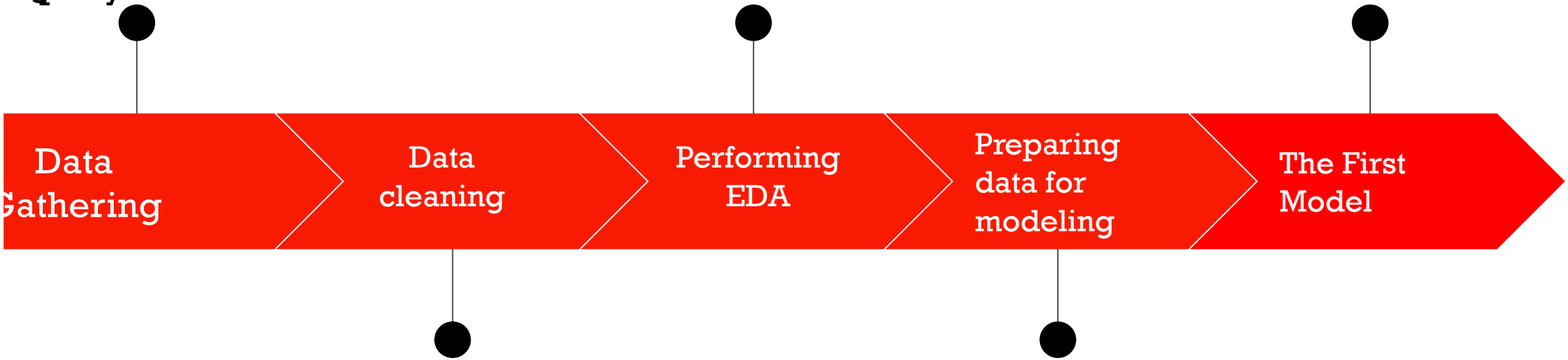
Increase conversion

Since we focused on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.

Loading & Observing the past data provided by the Company

Univariate, Bivariate and Heatmap for numerical and Categorical columns

Performing pre-requisites for RFE and Logistic Regression



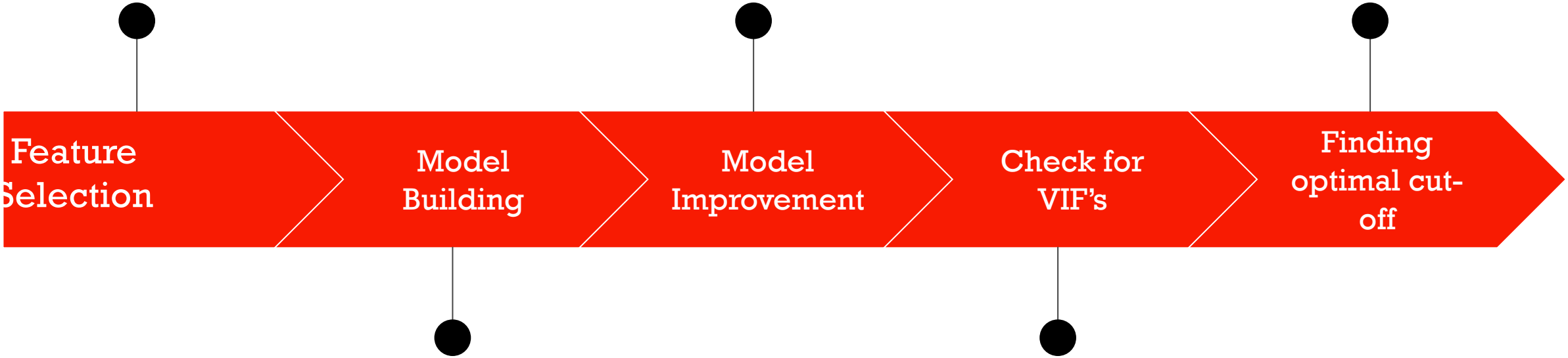
Duplicate removal, null value treatment, unnecessary column elimination, etc.

Outlier Treatment, Feature-Standardization

Selection of top 10 features using RFE

Reduction of columns and Model re-building

Verifying our Final Model Accuracy etc. with model built with PCA

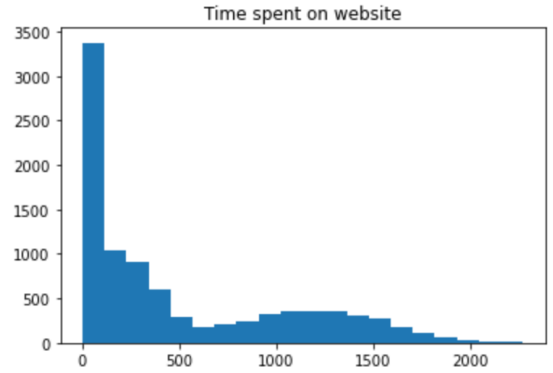


Model building using RFE for selected columns

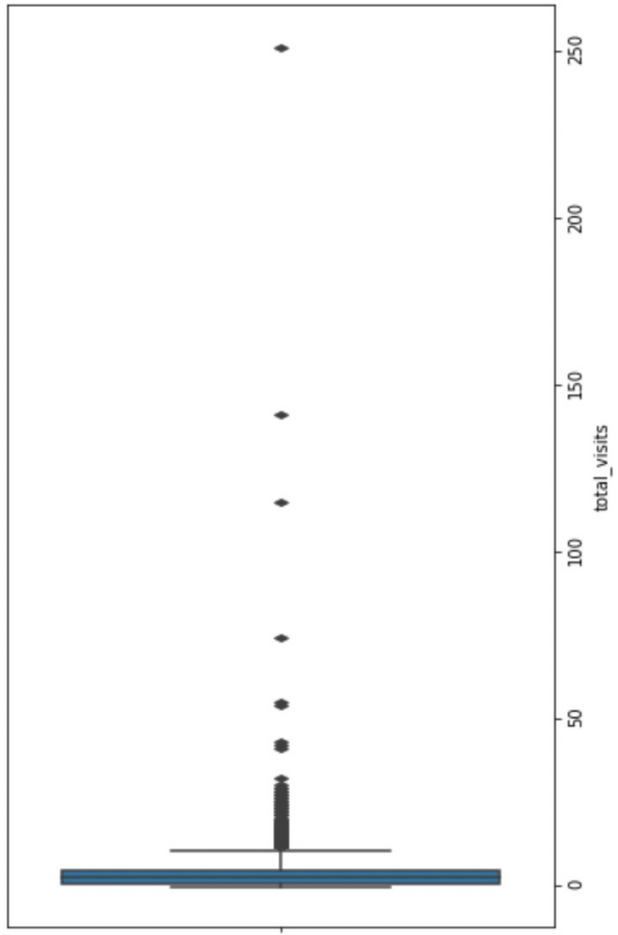
Final Model Analysis and performance on Test Data

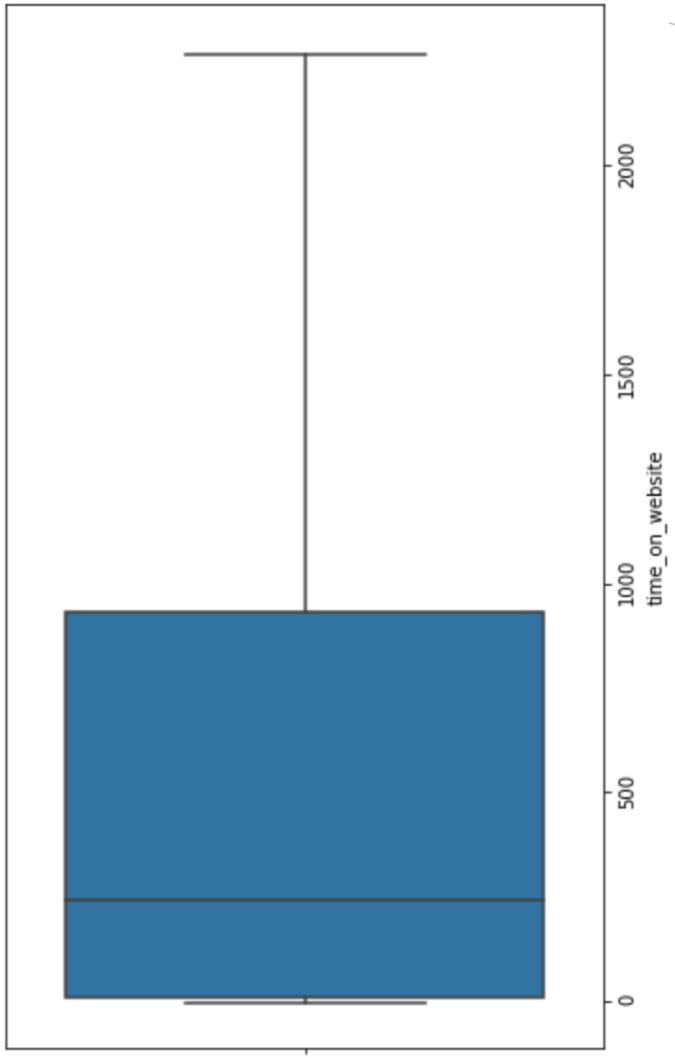
Total Visits

Inferences through Visualisation



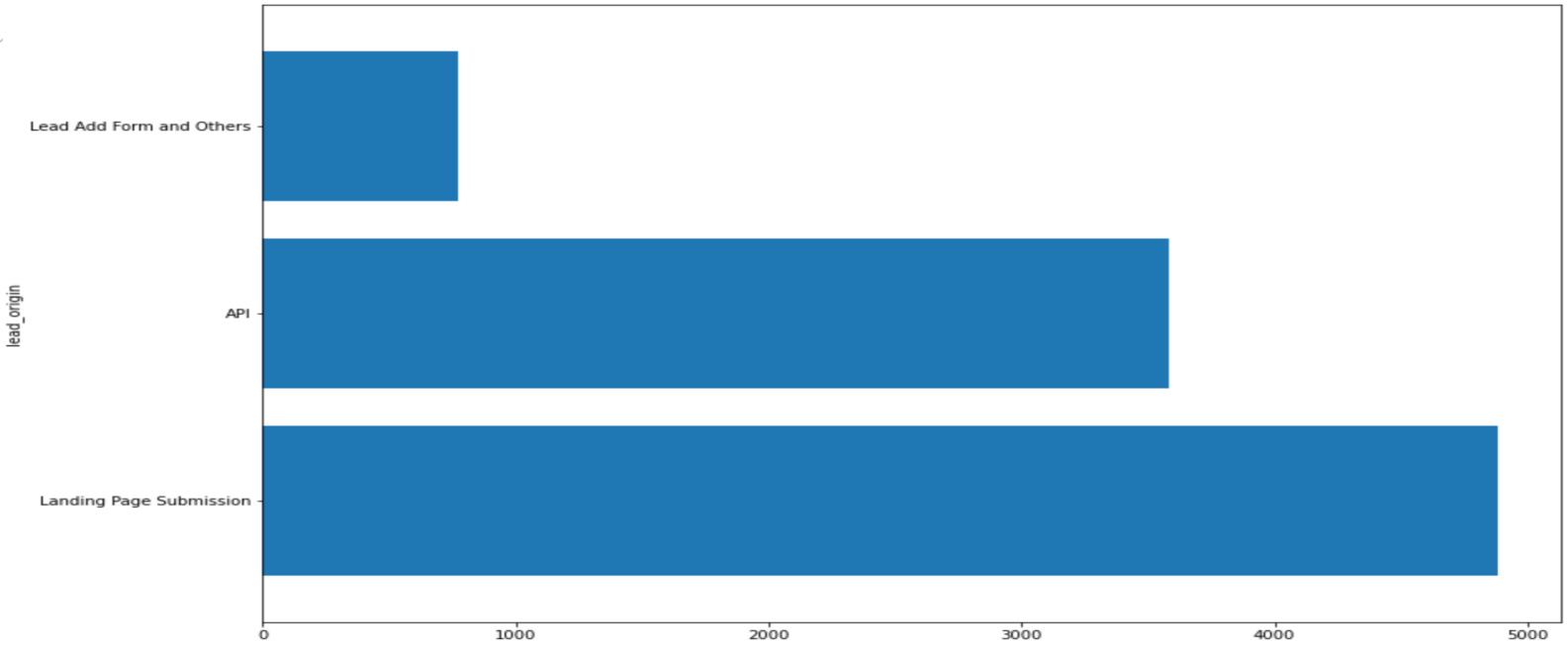
The peak seems unusually high, we will check for outliers in similar columns.

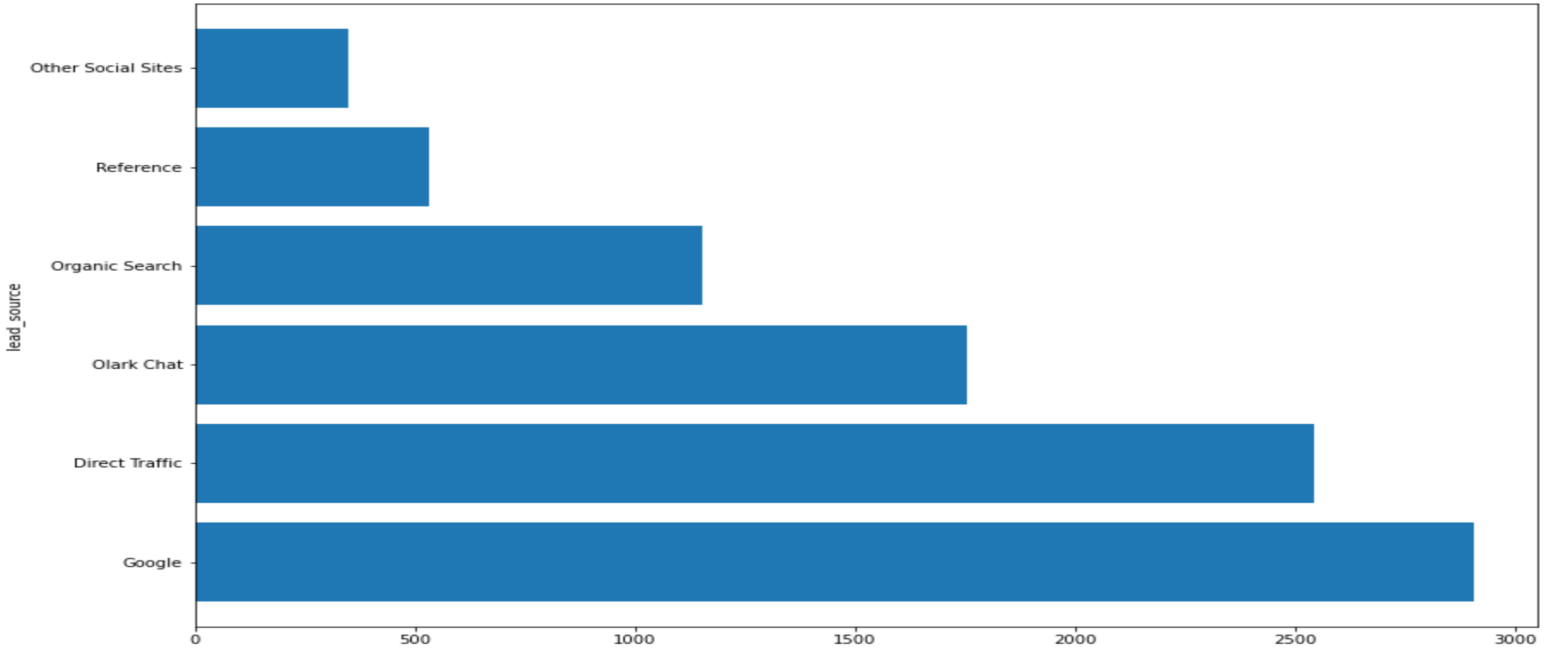




Total Leads

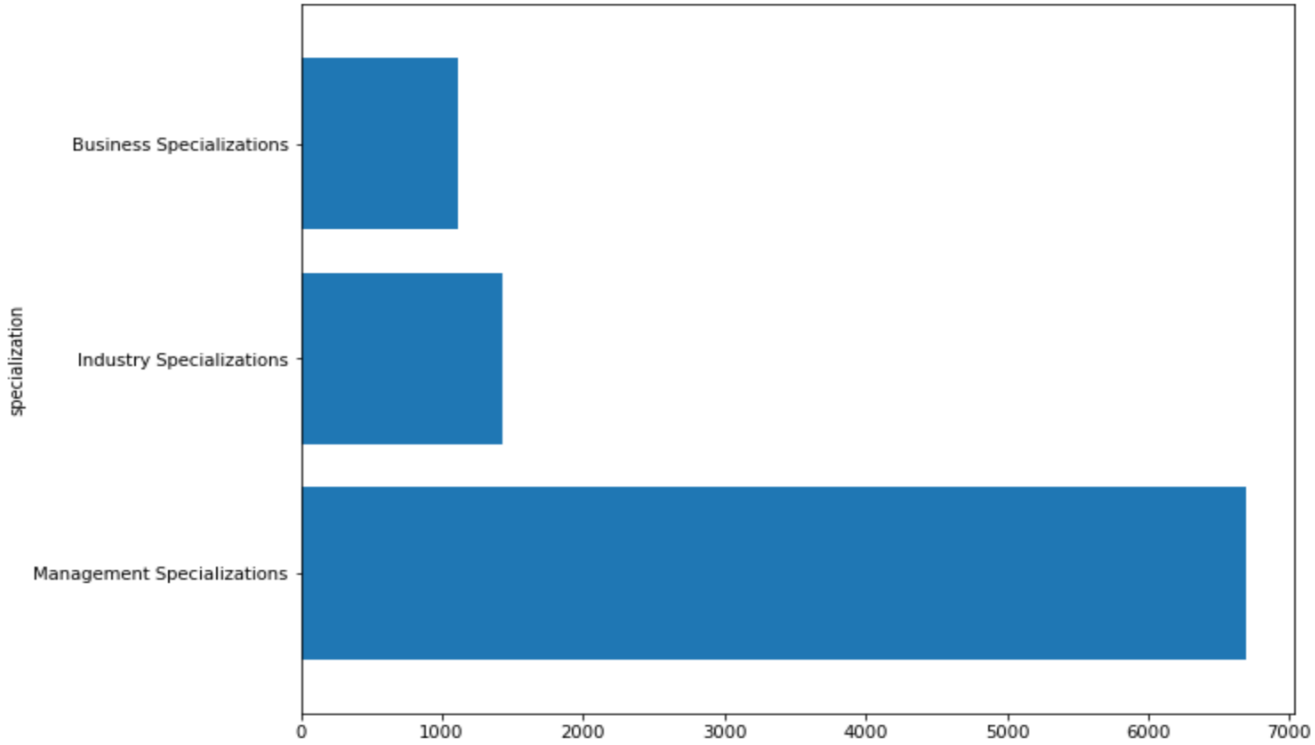
Inferences through Visualisation

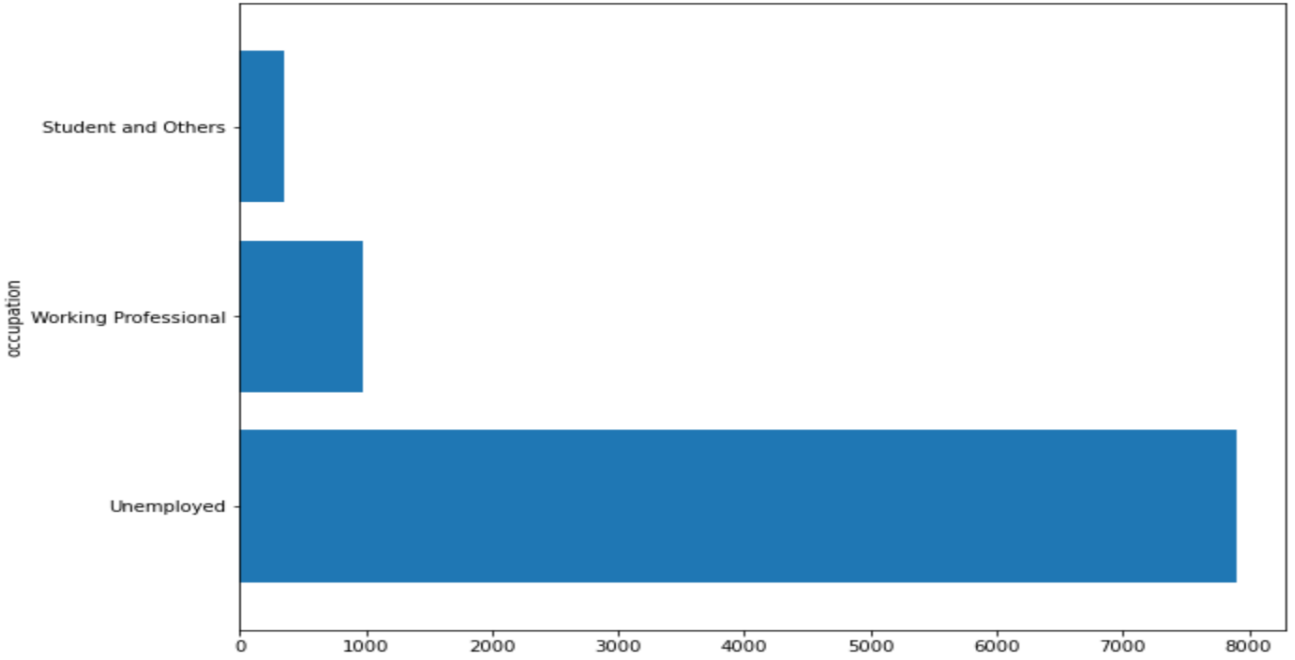




Total Leads

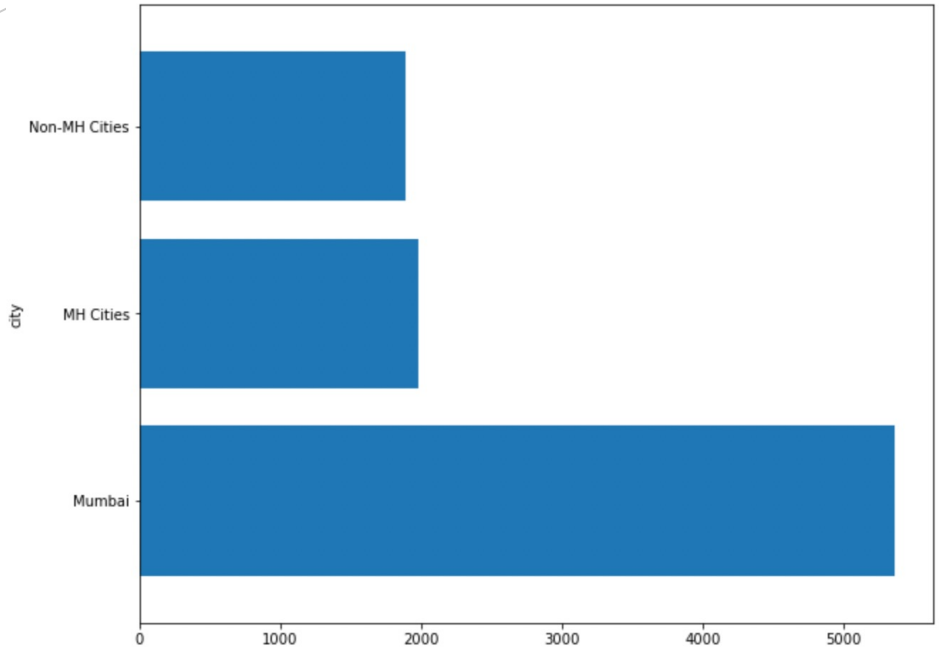
Inferences through Visualisation

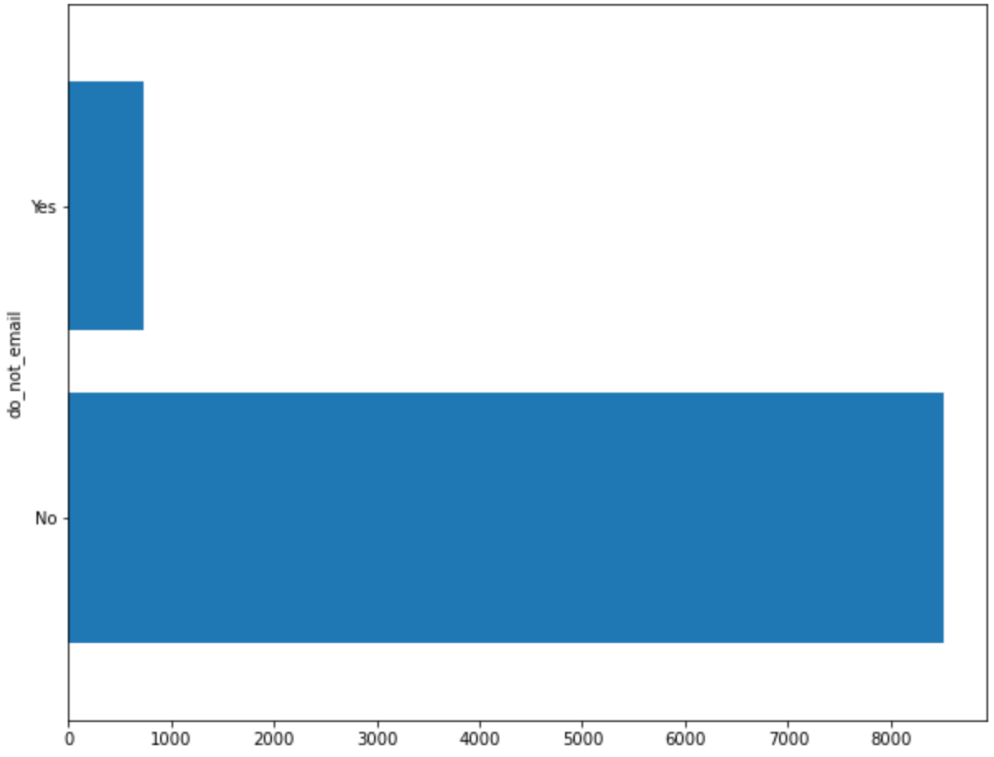


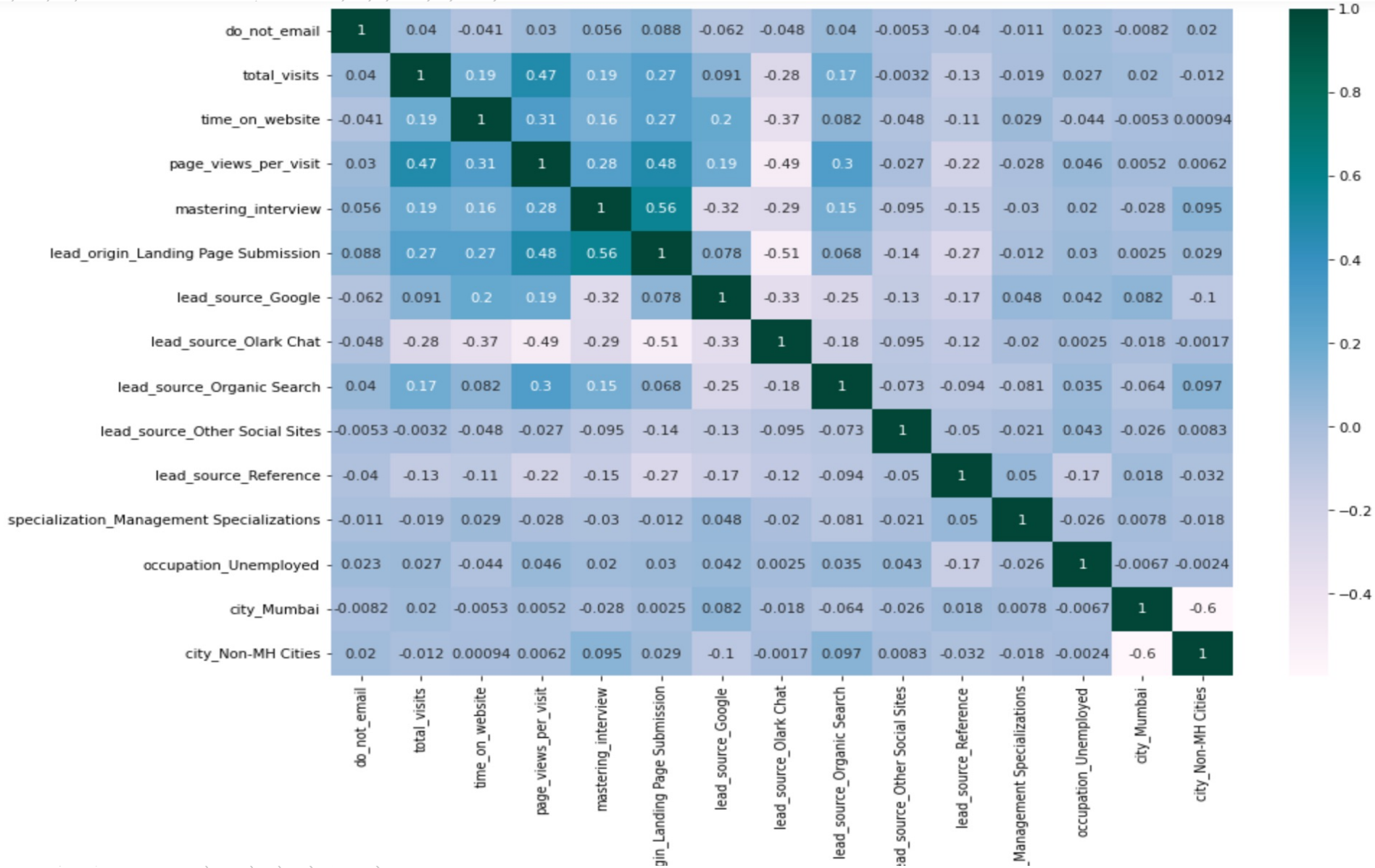


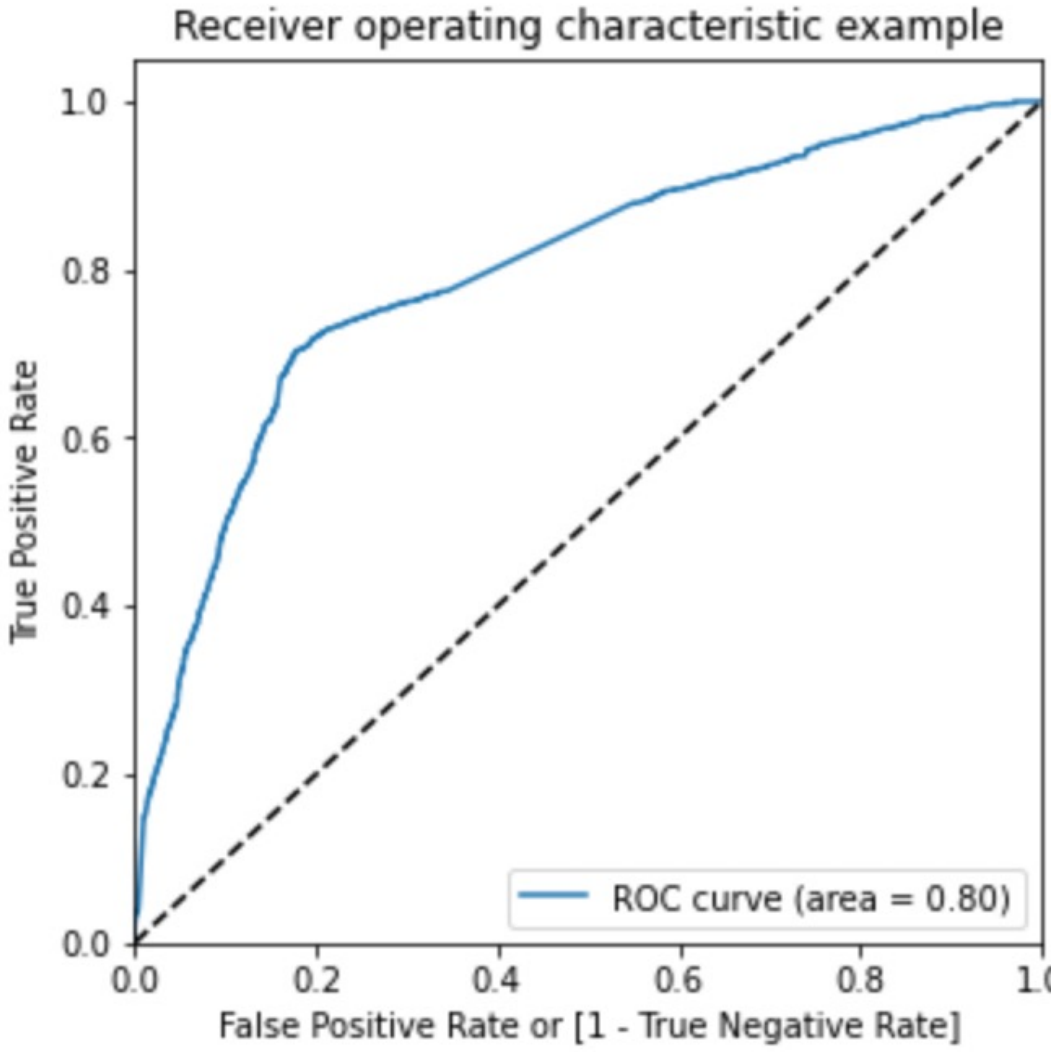
Total Leads

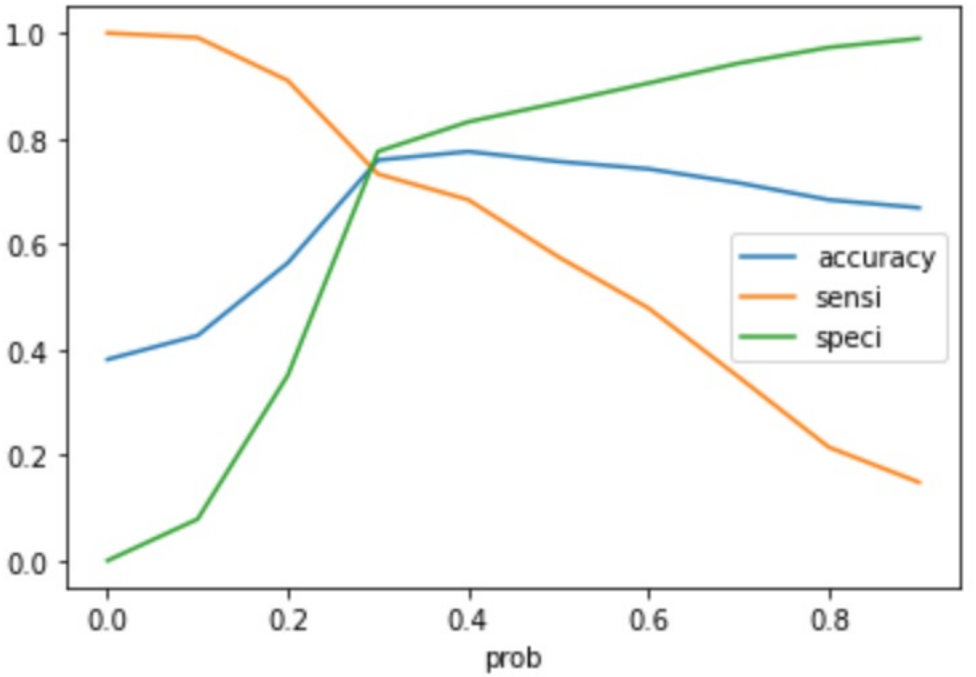
Inferences through Visualisation

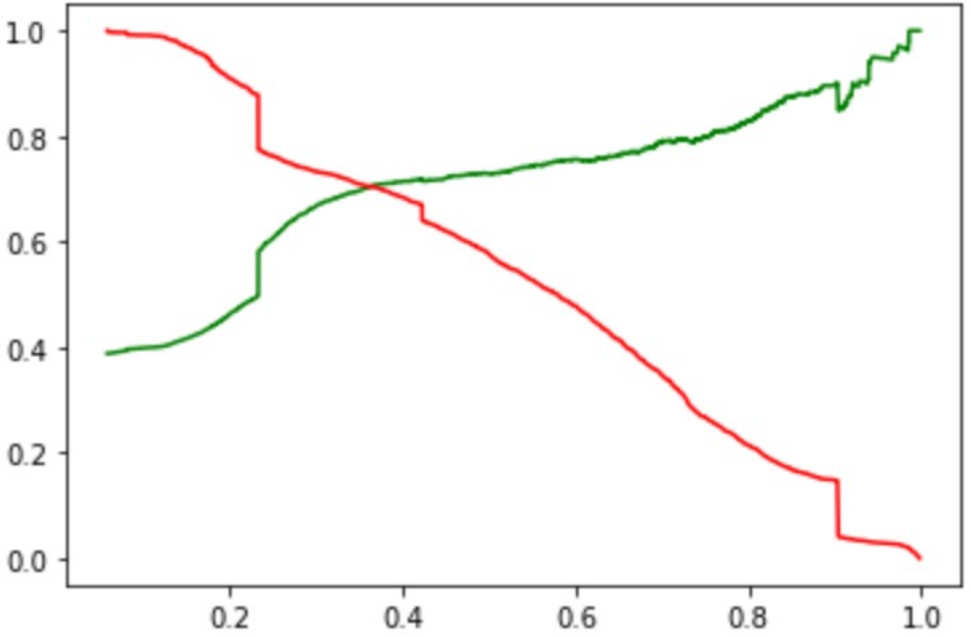





do_not_email 1 0.04 -0.041 0.03 0.056 0.088 -0.062 -0.048 0.04 -0.0053 -0.04 -0.011 0.023 -0.0082 0.02
total_visits 0.04 1 0.19 0.47 0.19 0.27 0.091 -0.28 0.17 -0.0032 -0.13 -0.019 0.027 0.02 -0.012
time_on_website -0.041 0.19 1 0.31 0.16 0.27 0.2 -0.37 0.082 -0.048 -0.11 0.029 -0.044 -0.0053 0.00094
page_views_per_visit 0.03 0.47 0.31 1 0.28 0.48 0.19 -0.49 0.3 -0.027 -0.22 -0.028 0.046 0.0052 0.0062
mastering_interview 0.056 0.19 0.16 0.28 1 0.56 -0.32 -0.29 0.15 -0.095 -0.15 -0.03 0.02 -0.028 0.095
lead_origin_Landing Page Submission 0.088 0.27 0.27 0.48 0.56 1 0.078 -0.51 0.068 -0.14 -0.27 -0.012 0.03 0.0025 0.029
lead_source_Google -0.062 0.091 0.2 0.19 -0.32 0.078 1 -0.33 -0.25 -0.13 -0.17 0.048 0.042 0.082 -0.1
lead_source_Olark Chat -0.048 -0.28 -0.37 -0.49 -0.29 -0.51 -0.33 1 -0.18 -0.095 -0.12 -0.02 0.0025 -0.018 -0.0017
lead_source_Organic Search 0.04 0.17 0.082 0.3 0.15 0.068 -0.25 -0.18 1 -0.073 -0.094 -0.081 0.035 -0.064 0.097
lead_source_Other Social Sites -0.0053 -0.0032 -0.048 -0.027 -0.095 -0.14 -0.13 -0.095 -0.073 1 -0.05 -0.021 0.043 -0.026 0.0083
lead_source_Reference -0.04 -0.13 -0.11 -0.22 -0.15 -0.27 -0.17 -0.12 -0.094 -0.05 1 0.05 -0.17 0.018 -0.032
specialization_Management Specializations -0.011 -0.019 0.029 -0.028 -0.03 -0.012 0.048 -0.02 -0.081 -0.021 0.05 1 -0.026 0.0078 -0.018
occupation_Unemployed 0.023 0.027 -0.044 0.046 0.02 0.03 0.042 0.0025 0.035 0.043 -0.17 -0.026 1 -0.0067 -0.0024
city_Mumbai -0.0082 0.02 -0.0053 0.0052 -0.028 0.0025 0.082 -0.018 -0.064 -0.026 0.018 0.0078 -0.0067 1 -0.6
city_Non-MH Cities 0.02 -0.012 0.00094 0.0062 0.095 0.029 -0.1 -0.0017 0.097 0.0083 -0.032 -0.018 -0.0024 -0.6 1
do_not_email
total_visits
time_on_website
page_views_per_visit
mastering_interview
gin_Landing Page Submission
lead_source_Google
lead_source_Olark Chat
lead_source_Organic Search
ead_source_Other Social Sites
lead_source_Reference
_Management Specializations
occupation_Unemployed
city_Mumbai
city_Non-MH Cities
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4







Linear Regression Final Model Parameters
Area under ROC = 0.80
Intermediate cut-off = 0.35
Final cut-off = 0.40

Model Analysis

- Overall accuracy on Test set: 0.77%
- Sensitivity of our logistic regression model: 68%
- Specificity of our logistic regression model: 83%

Business insights from the model

Top 3 variables in model, that contribute towards lead conversion are:

- Total Time Spent on Website.
- Page views per visit.
- Lead source Google

Business insights from the model

Top 3 variables in my model, that should be focused are:

- Time on website (positively impacting)
- Lead Source Olark Chat (Positively impacting)
- Lead Source Google (Positively impacting)

Logistic Regression model

Model Summary

Our Logistic Regression Model is decent and accurate enough, when compared to the model derived using PCA.

- Accuracy = 77%
- Sensitivity = 68%
- Specificity = 83%

We can vary these parameters by varying the cut-off value and thus predict Hot leads based on scenarios like availability of extra resources and vice-versa.

Recommendations

X EDUCATION

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on their website since this helps in higher conversion.
- Get Total visits increased by advertising etc. since this helps in higher conversion(Main emphasis on google)
- More emphasis on the Olark Chat service since this is affecting the conversion quite a bit.
- Technically, we can generate this new set of leads by altering (moving down) the value of cut off (0.4) so as to include leads from our Logistic Regression Model.