X EDUCATION – LEAD SCORING CASE STUDY

Background

X Education Company

- X Education , An education company named sells online courses to industry professionals
- Many interested professionals land on their website
- The company markets its courses on several websites 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

Background

X Education Company

- **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 Company's Problem

- **X Education gets a lot of leads but 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'**

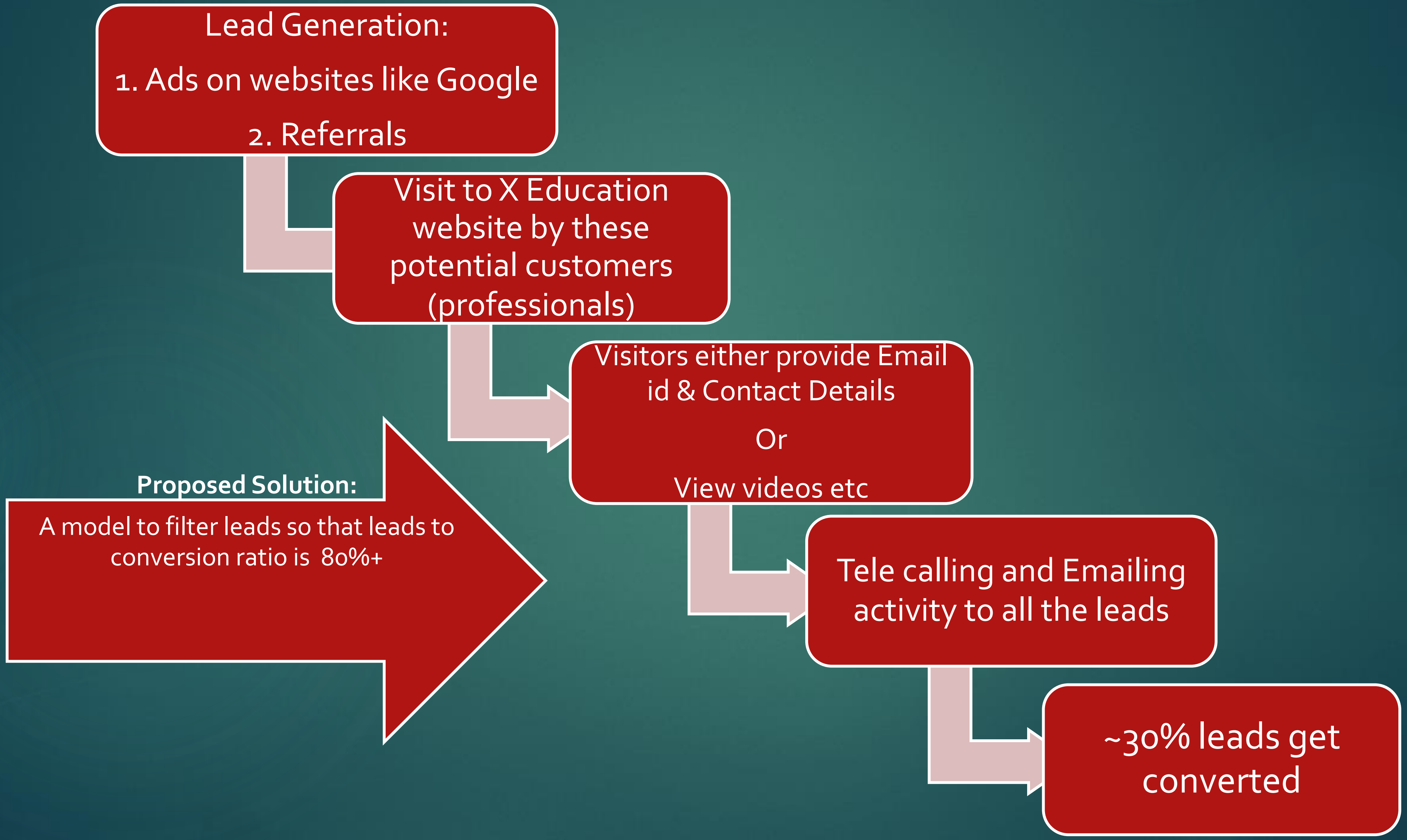
- **If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone**

Problem Statement

X Education Company’s Problem

- We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

LEAD – CONVERSION PROCESS



Lead Generation:

1. Ads on websites like Google

2. Referrals

Visit to X Education website by these potential customers (professionals)

Visitors either provide Email id & Contact Details

Or

View videos etc

Tele calling and Emailing activity to all the leads

~30% leads get converted

Proposed Solution:

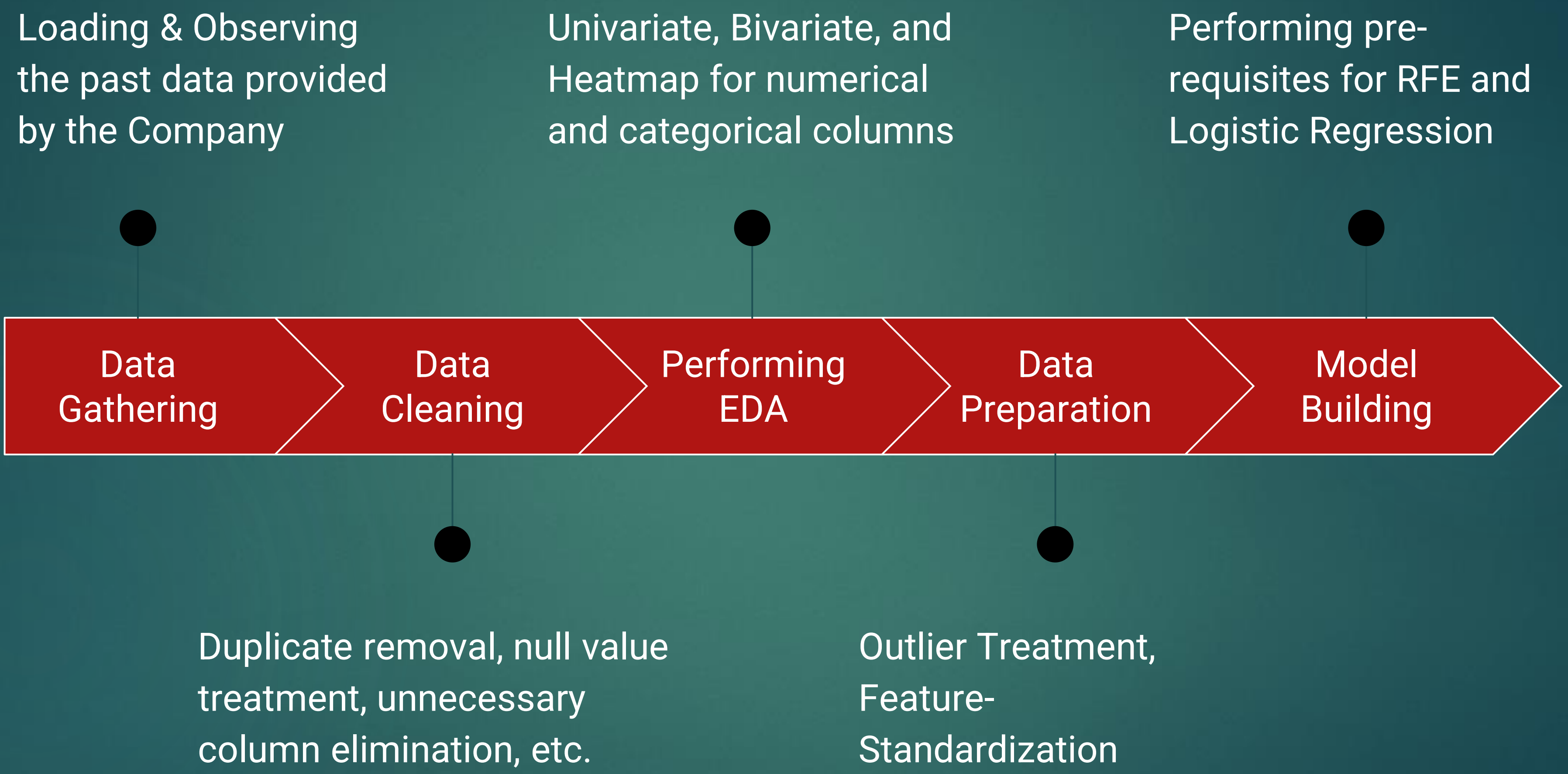
A model to filter leads so that leads to conversion ratio is 80%+

Solution

Selection of Hot Leads

- For our Problem Solution, the crucial part is to accurately identify hot leads.
- The more accurate we obtain the hot lead, the more chance we get of higher conversion ratio.
- Since we have a target of 80% conversion rate, we would want to obtain a high accuracy in obtaining hot leads.

IMPLEMENTATION



Data Gathering

Loading & Observing the past data provided by the Company

Data Cleaning

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

Performing EDA

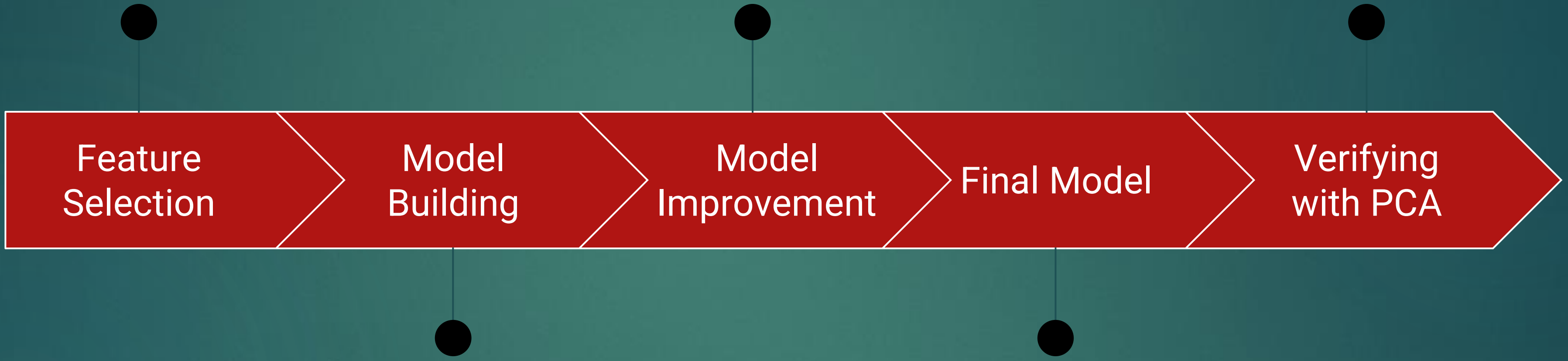
Univariate, Bivariate, and Heatmap for numerical and categorical columns

Data Preparation

Outlier Treatment, Feature-Standardization

Model Building

Performing pre-requisites for RFE and Logistic Regression



Feature Selection

Selection of top 25 features using RFE

Model Building

Model building using RFE for selected columns

Model Improvement

Reduction of columns and Model re-building

Final Model

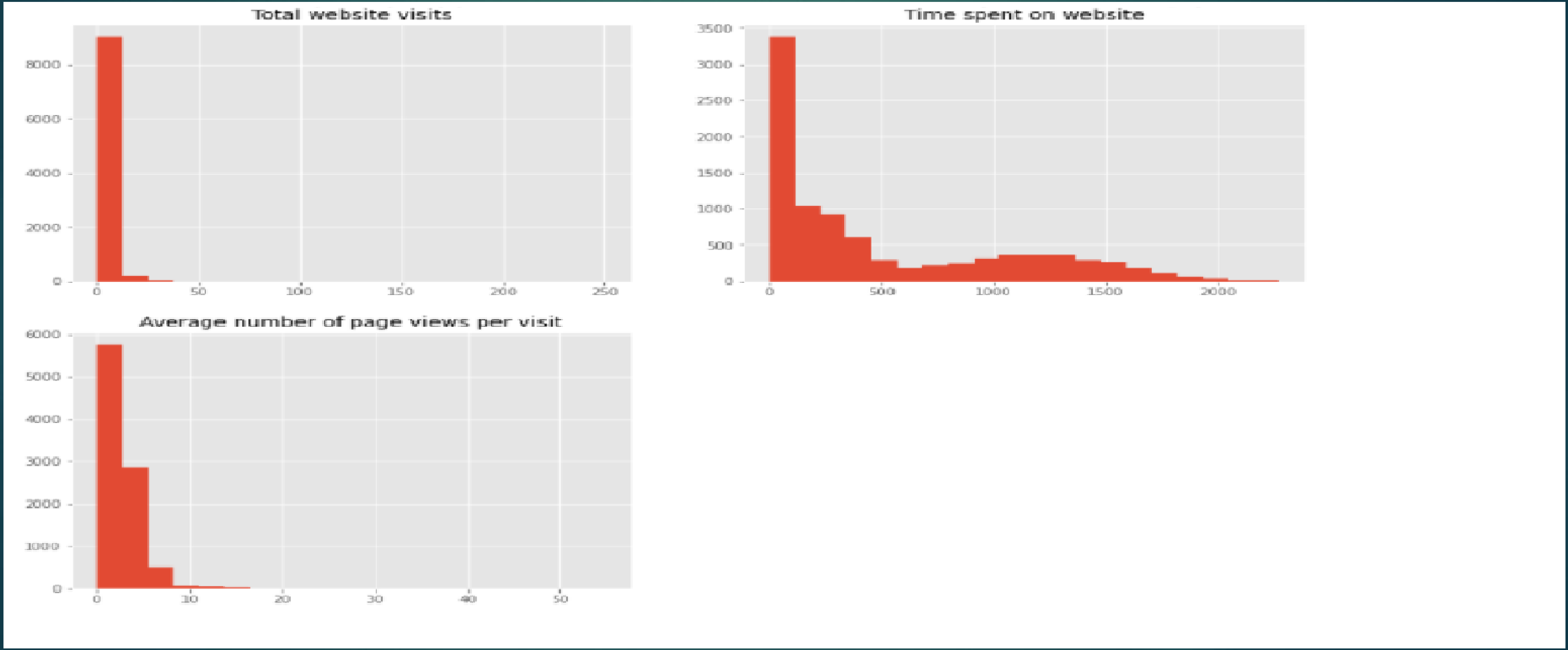
Final Model Analysis and performance on Test Data

Verifying with PCA

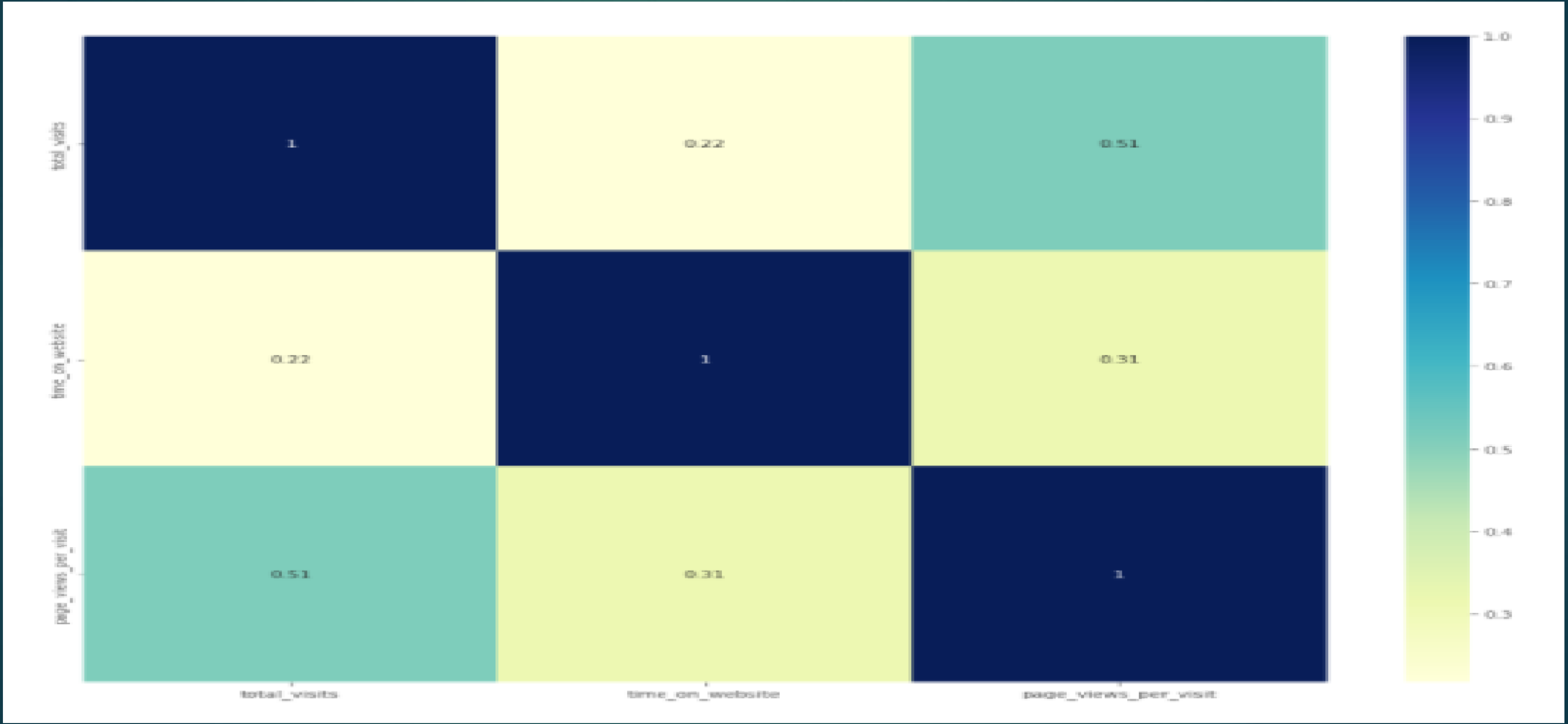
Verifying our Final Model Accuracy etc. with model built with PCA

VISUALIZATION

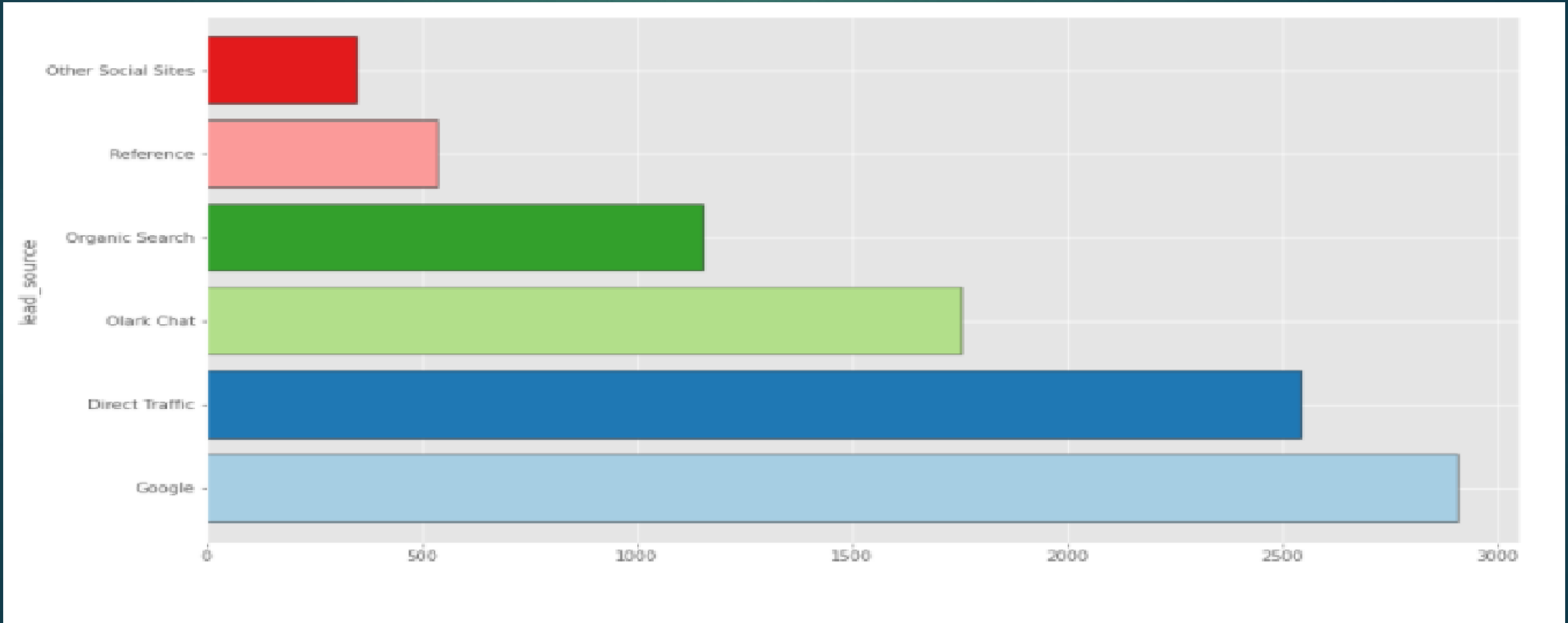
EDA plots depicting variation in numerical columns



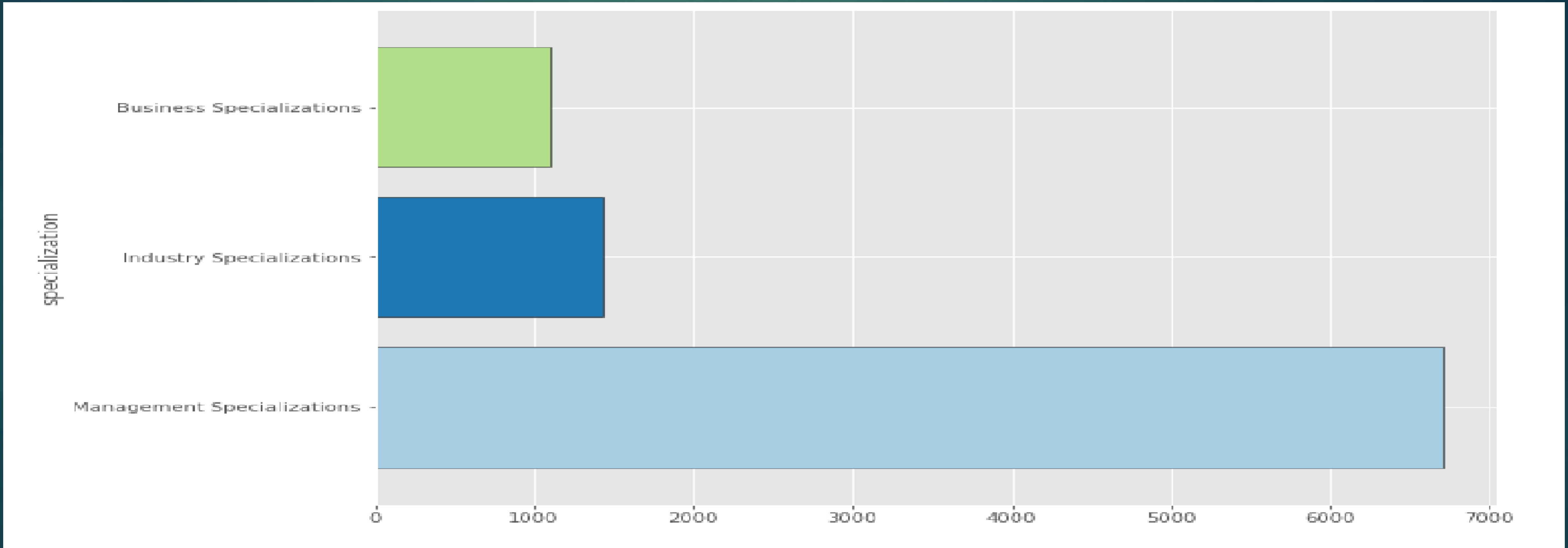
Correlation (Heat Map) of all selected numerical columns



EDA plot depicting variation in categorical column (Lead Source)

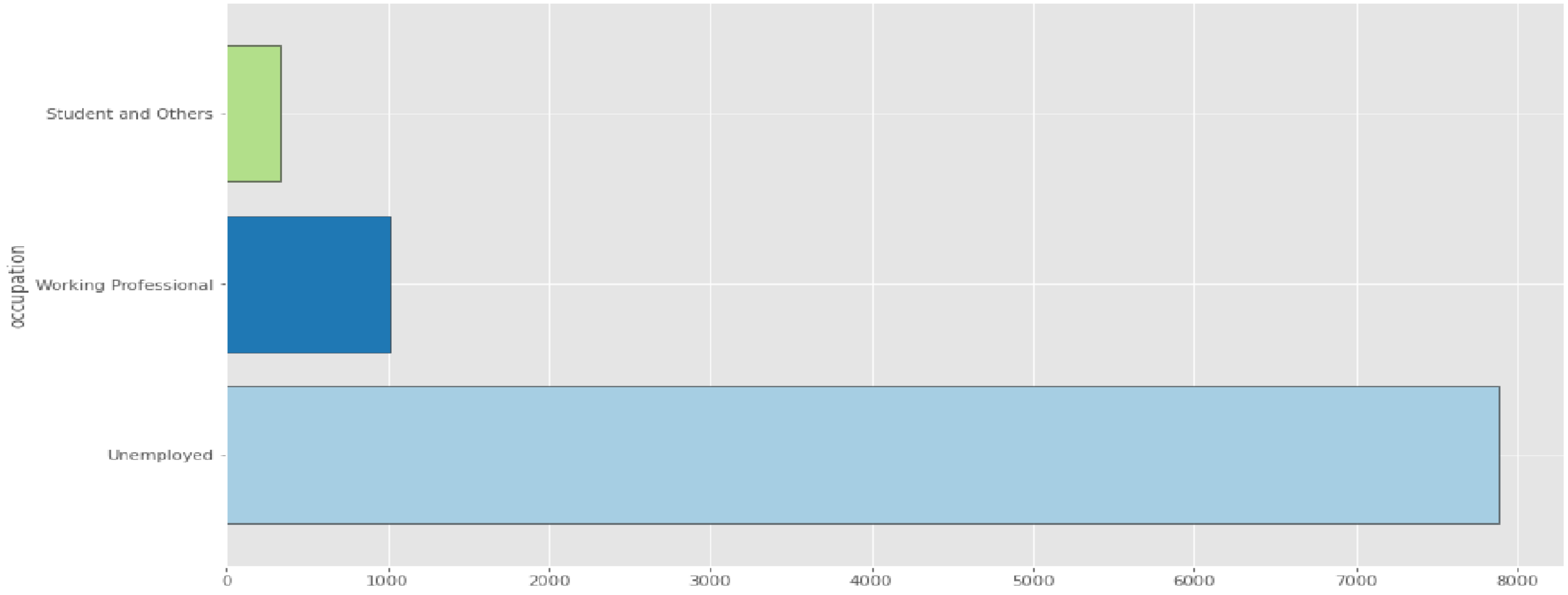


EDA plot depicting variation in categorical column (Specialization)



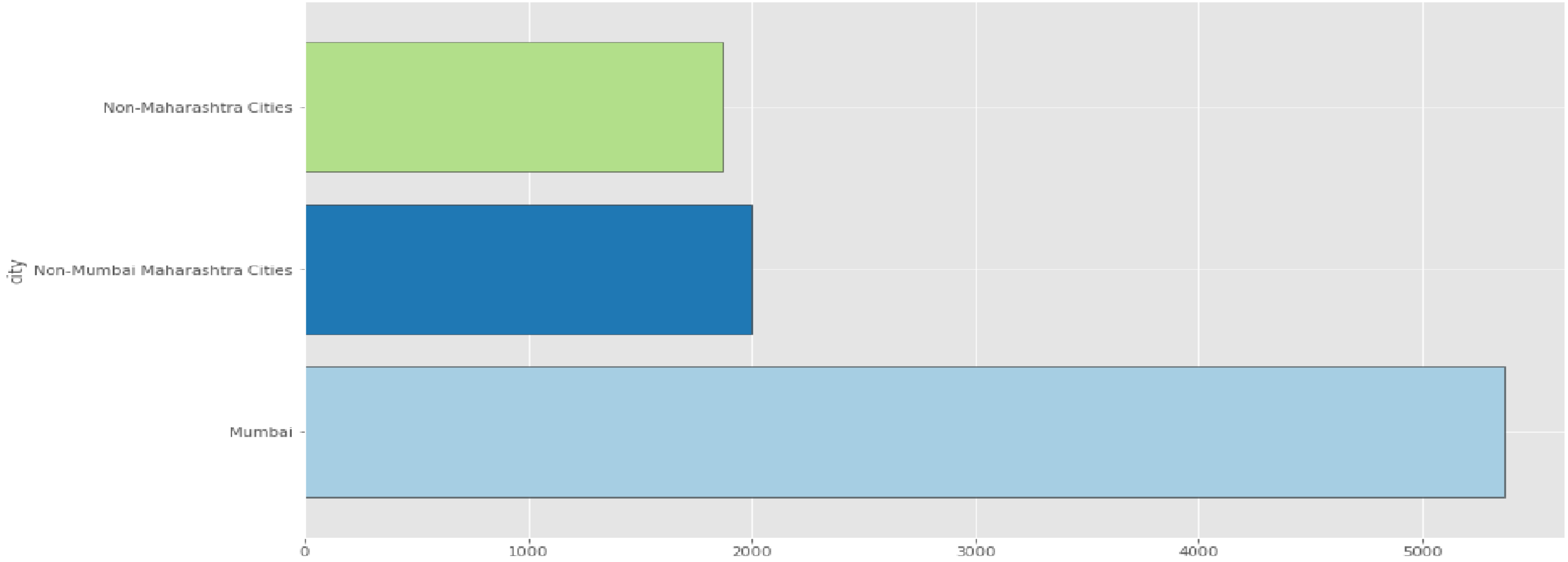
Most of the speciliazation taken are management

EDA plot depicting variation in categorical column (Occupation)



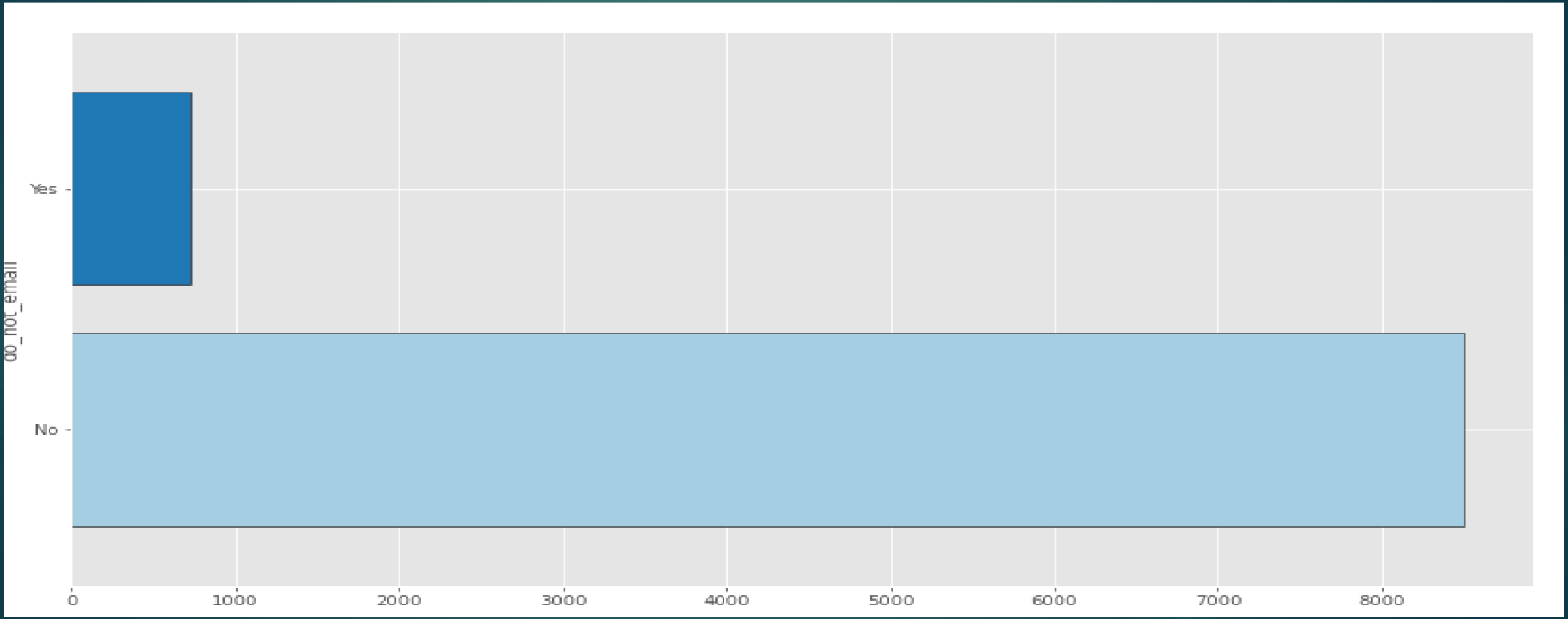
Unempployed users are the most significant leads

EDA plot depicting variation in categorical column (City)



Mumbai in particular and Maharashtra in general dominates the lead. This is likely due to the fact that the courses are based in Mumbai

EDA plot depicting variation in categorical column (Do not email)

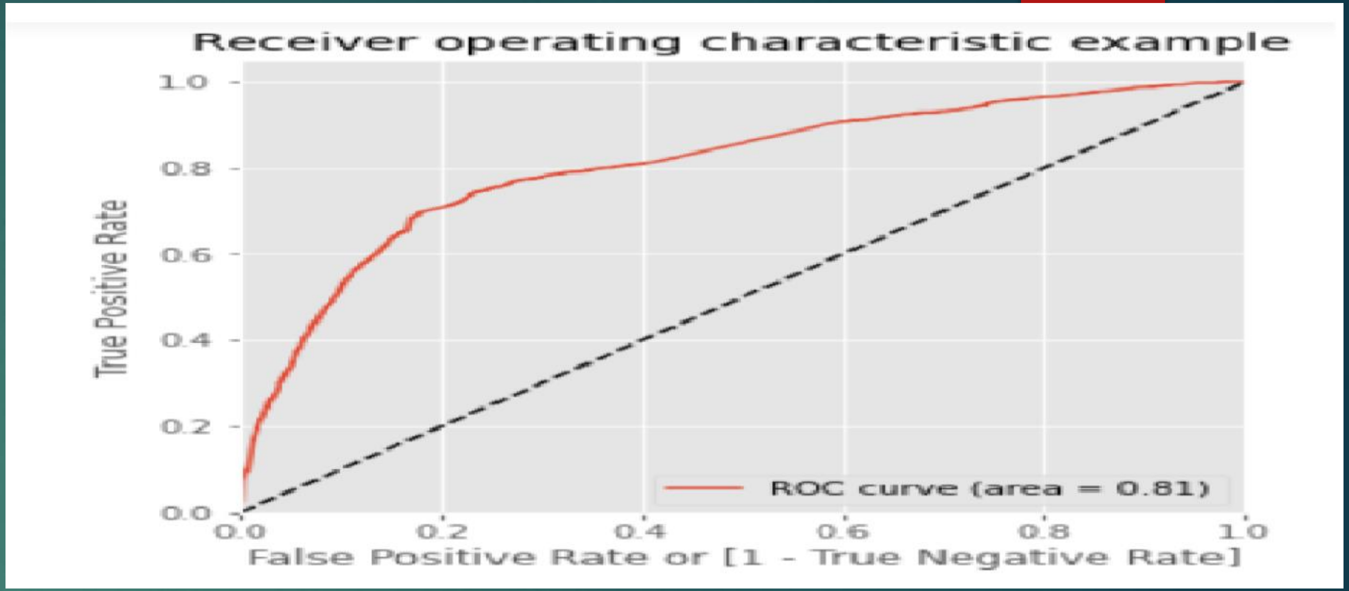


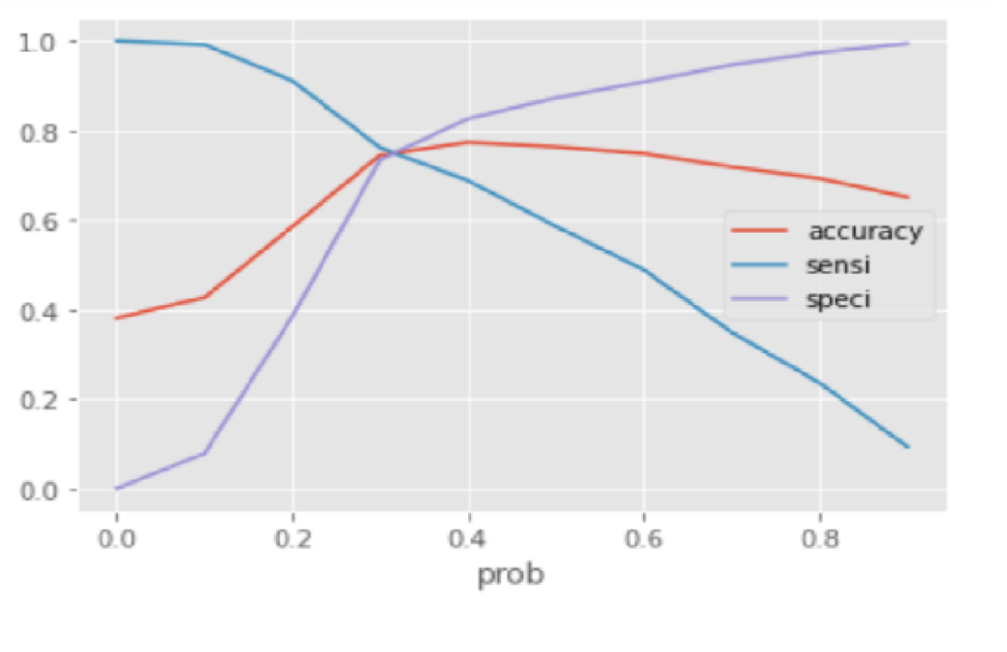
Linear Regression Final Model Parameters

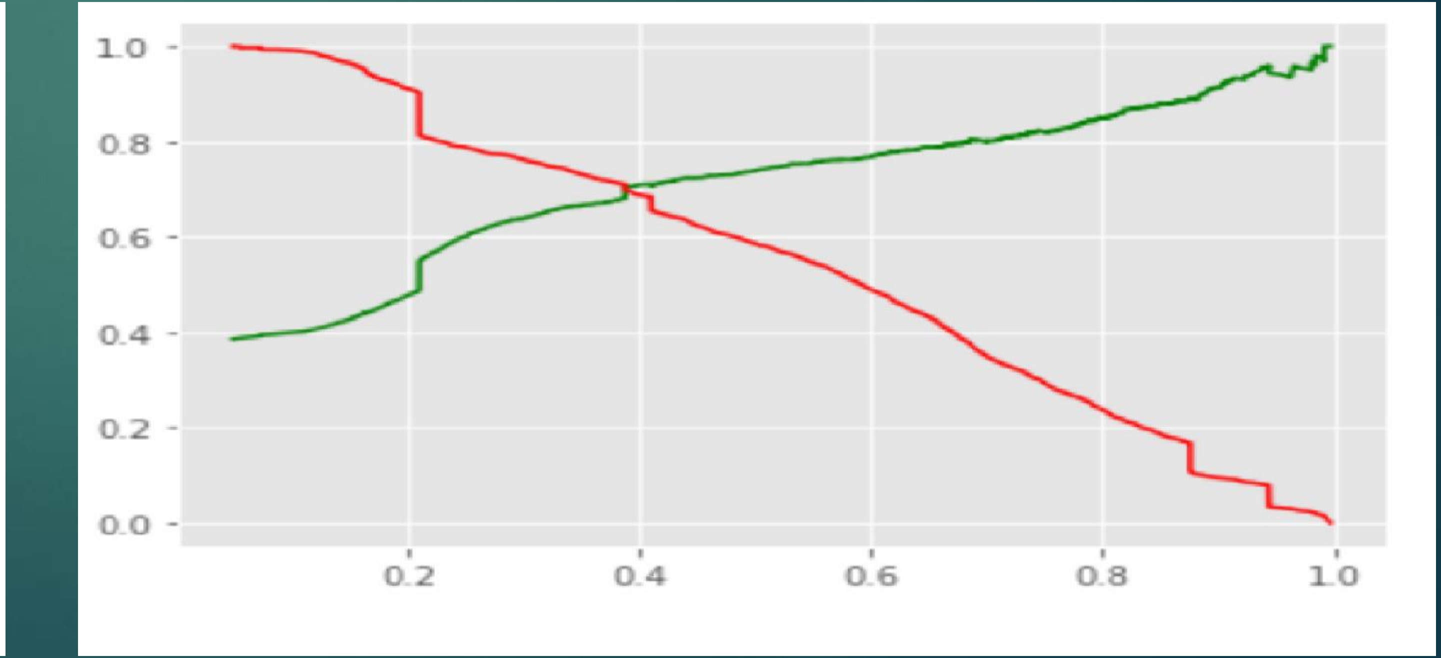
Area Under ROC = 0.81

Intermediate Cut-off = 0.3

Final Cut-off = 0.4







CONCLUSION

Model Analysis

Performance of our Final Model

- Overall accuracy on Test set: 74.60%

- Sensitivity of our logistic regression model: 76.52%

- Specificity of our logistic regression model: 73.73%

Inferences from Model

Business Insights Derived from our Model

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

- Total Time Spent on Website
- Do_not_Email
- TotalVisits

Conclusion 1 (LR Model)

Our Logistic Regression Model is decent and accurate enough, when compared to the model derived using PCA, with 74.6 % Accuracy on Test Set, 76.5 % Sensitivity and 73.3 % Specificity.
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.

Conclusion 2 (Recommendation)

- Target leads that spend a lot of time on X-Education site (Total Time Spent on Website)
- Target leads that repeatedly visit the site (Page Views Per Visit). However they might be repeatedly visiting to compare courses from the other sites, as the number of visits might be for that reason. So the interns should be a bit more aggressive and should ensure competitive points where X-Education is better, are strongly highlighted.
- Target leads that have come through References as they have a higher probability of converting
- Students can be approached, but they will have a lower probability of converting due to the course being industry based. However, this can also be a motivating factor to ensure industry readiness by the time they complete their education

THANK YOU