LEAD SCORING CASE STUDY FOR X EDUCATION

AIM: TO IDENTIFY THE MOST POTENTIAL LEADS, ALSO KNOWN AS 'HOT LEADS'

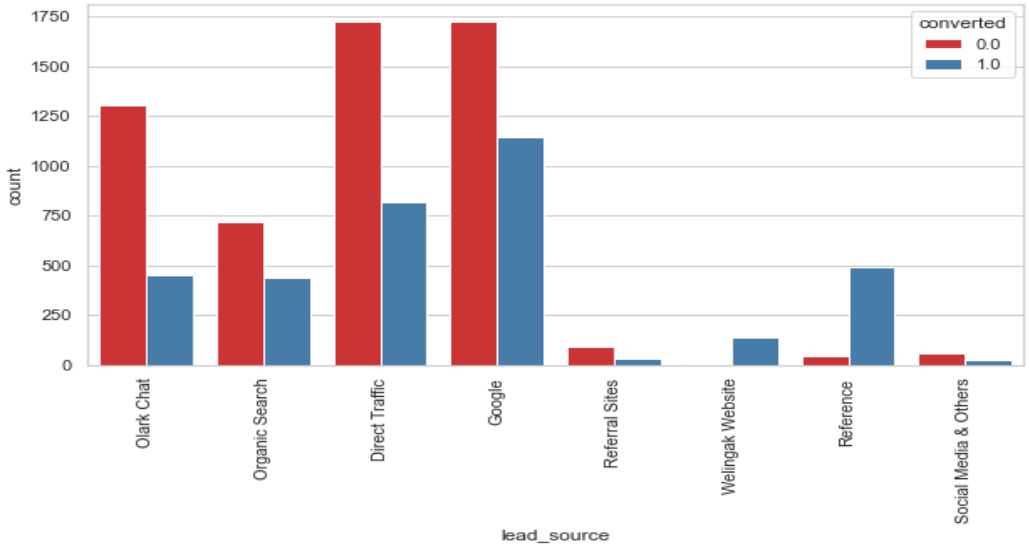
OBJECTIVE: WITH ABOUT 37 PREDICTOR VARIABLE PROVIDED WE HAD TO BUILD A LOGISTIC REGRESSION MODEL THAT WILL HELP THE SALES TEAM FOCUS THEIR ENERGY AND TIME TO ONLY THE HIGHLY COVERABLE CANDIDATES. THEREBY INCREASING THE CONVERSION RATE WHICH AT PRESENT IS ONLY 30%.

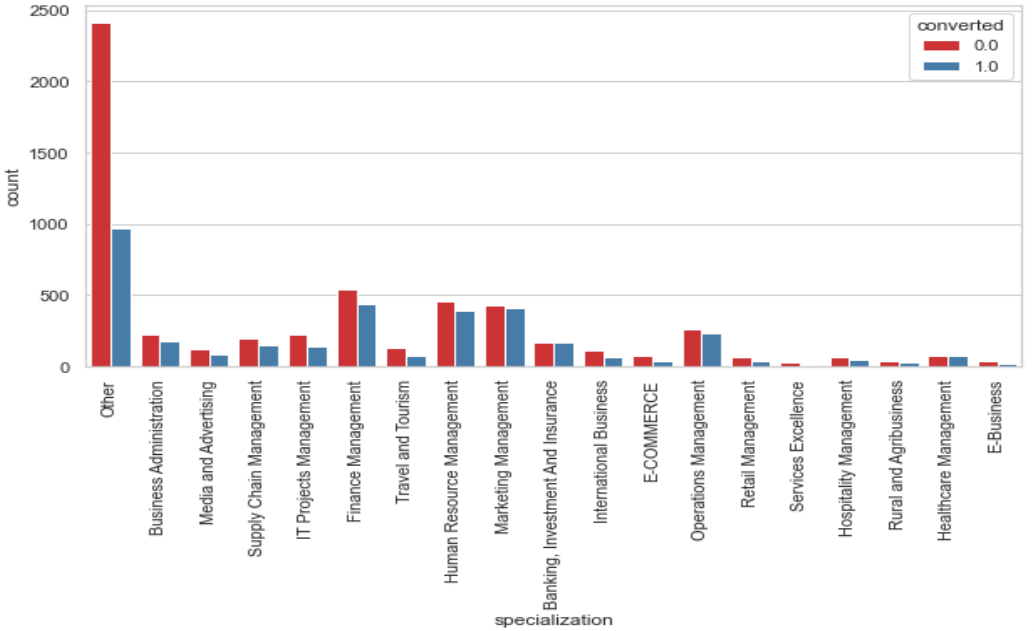
PROJECT BY: 1. KAMIL SHEIKH

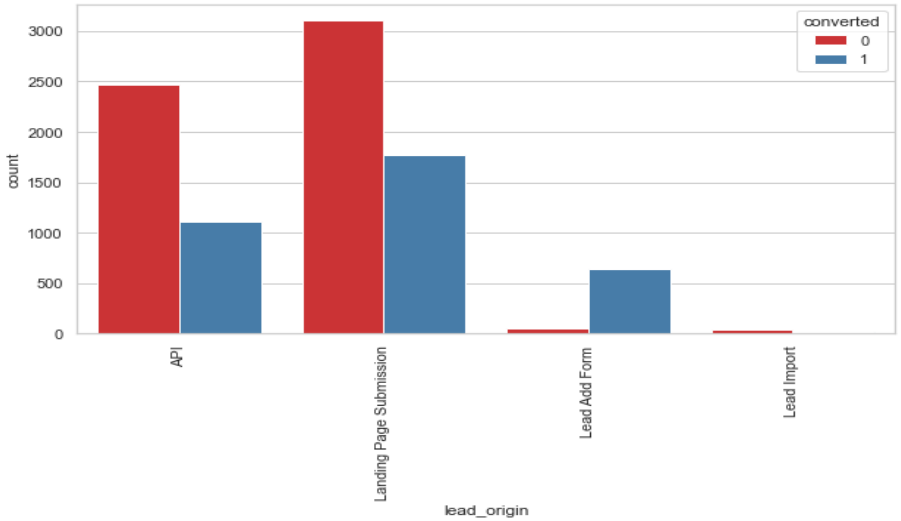
2. SNEHA SOMAN

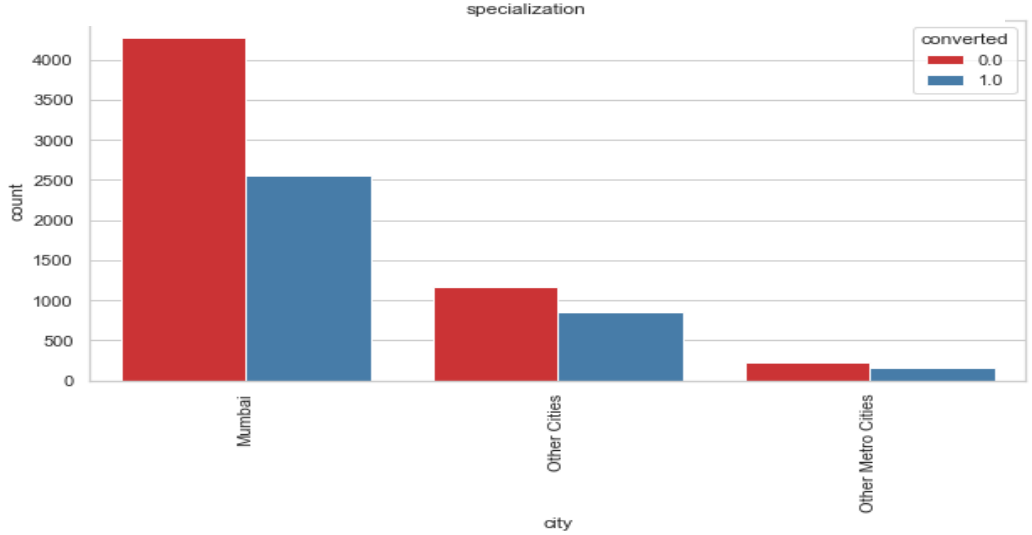
DATE: 19TH JANUARY 2022

RETAINED FEATURE VARIABLES AND THEIR RELATION WITH "CONVERTED"

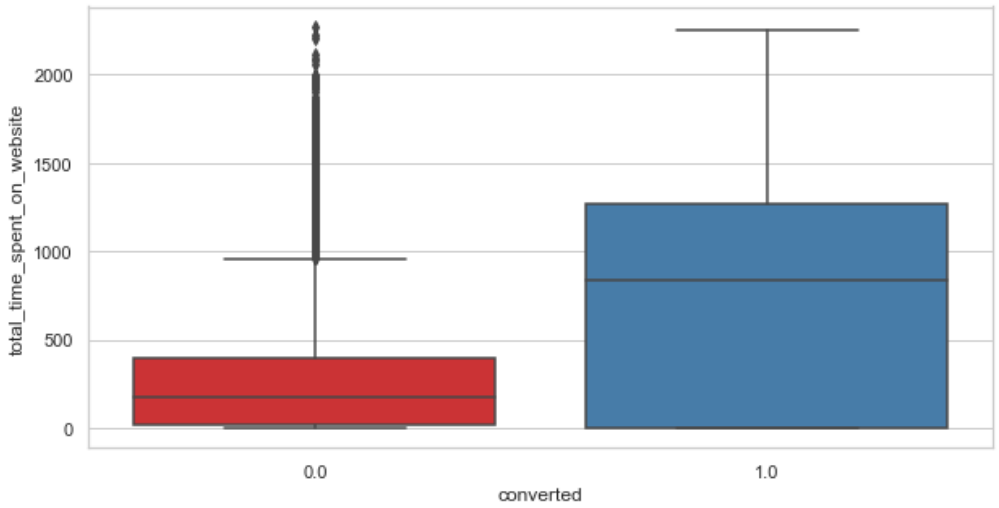


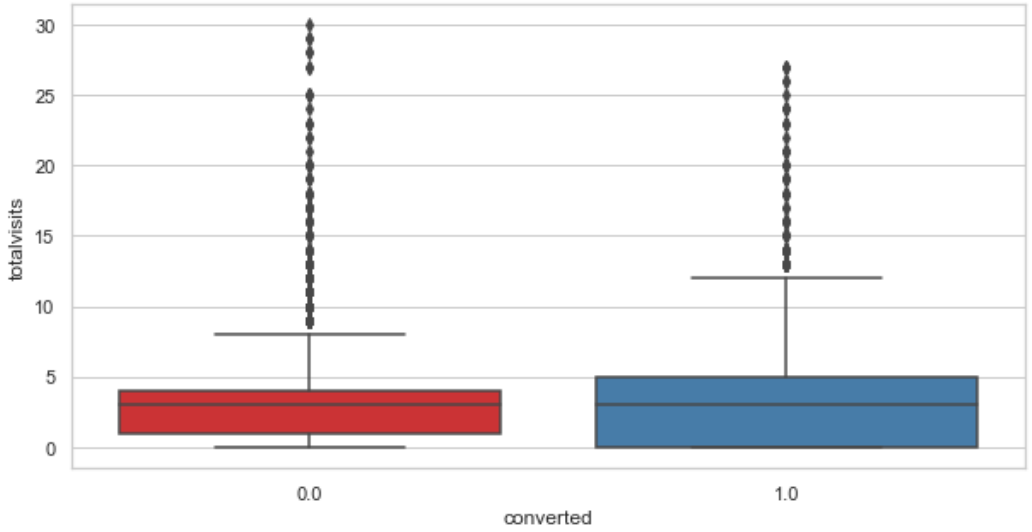


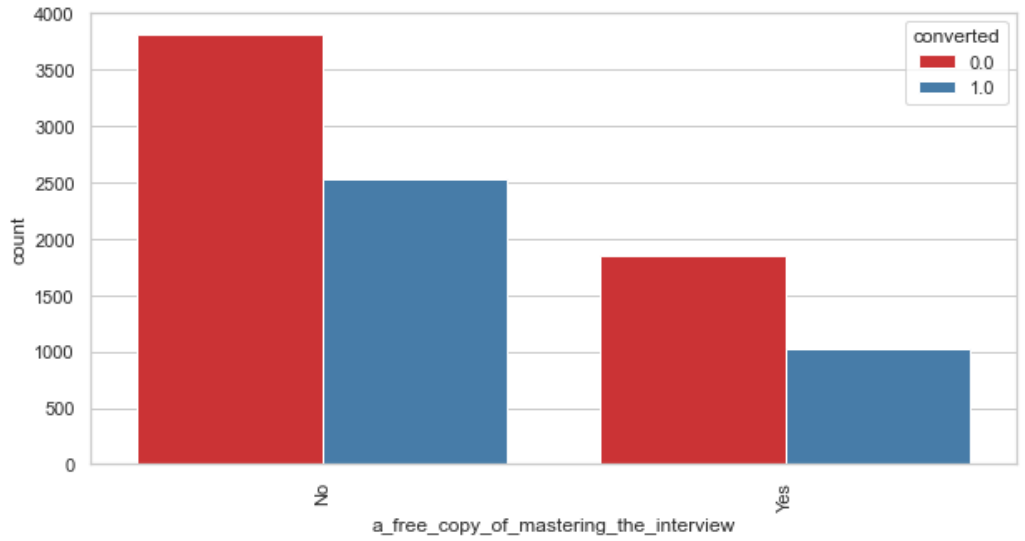


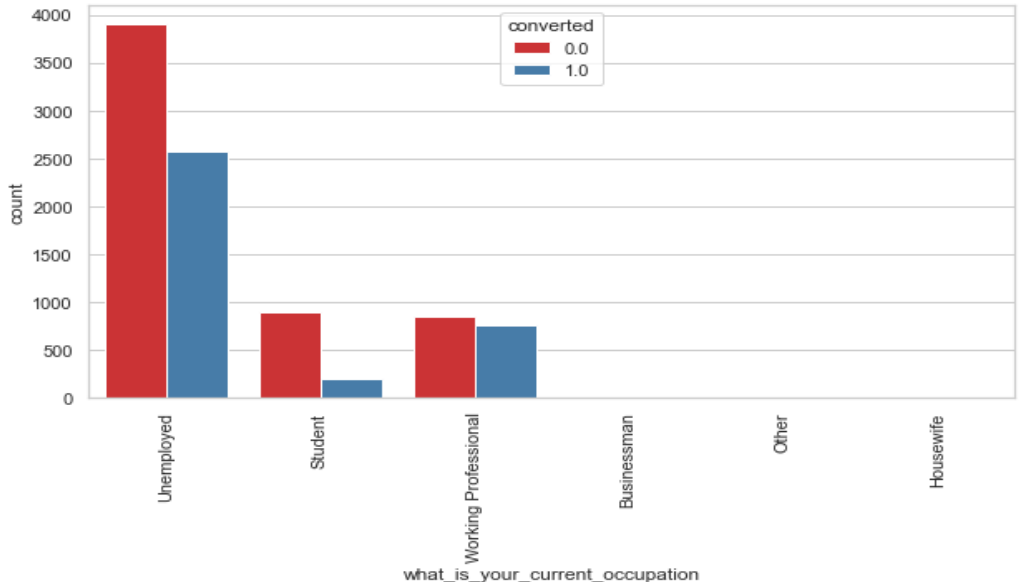


RETAINED FEATURE VARIABLES AND THEIR RELATION WITH "CONVERTED"









MAJOR HURDLES DURING DATA CLEANING

1) The Dataset had many missing values, most of them being categorical in nature.
2) There were major data discrepancies in Country and City
3) Most of the columns in the data set had very skewed data.
4) Some of the key variables used in the prediction model had outliers.

```
df1.country[df1.city=='Mumbai'].value_counts()
```

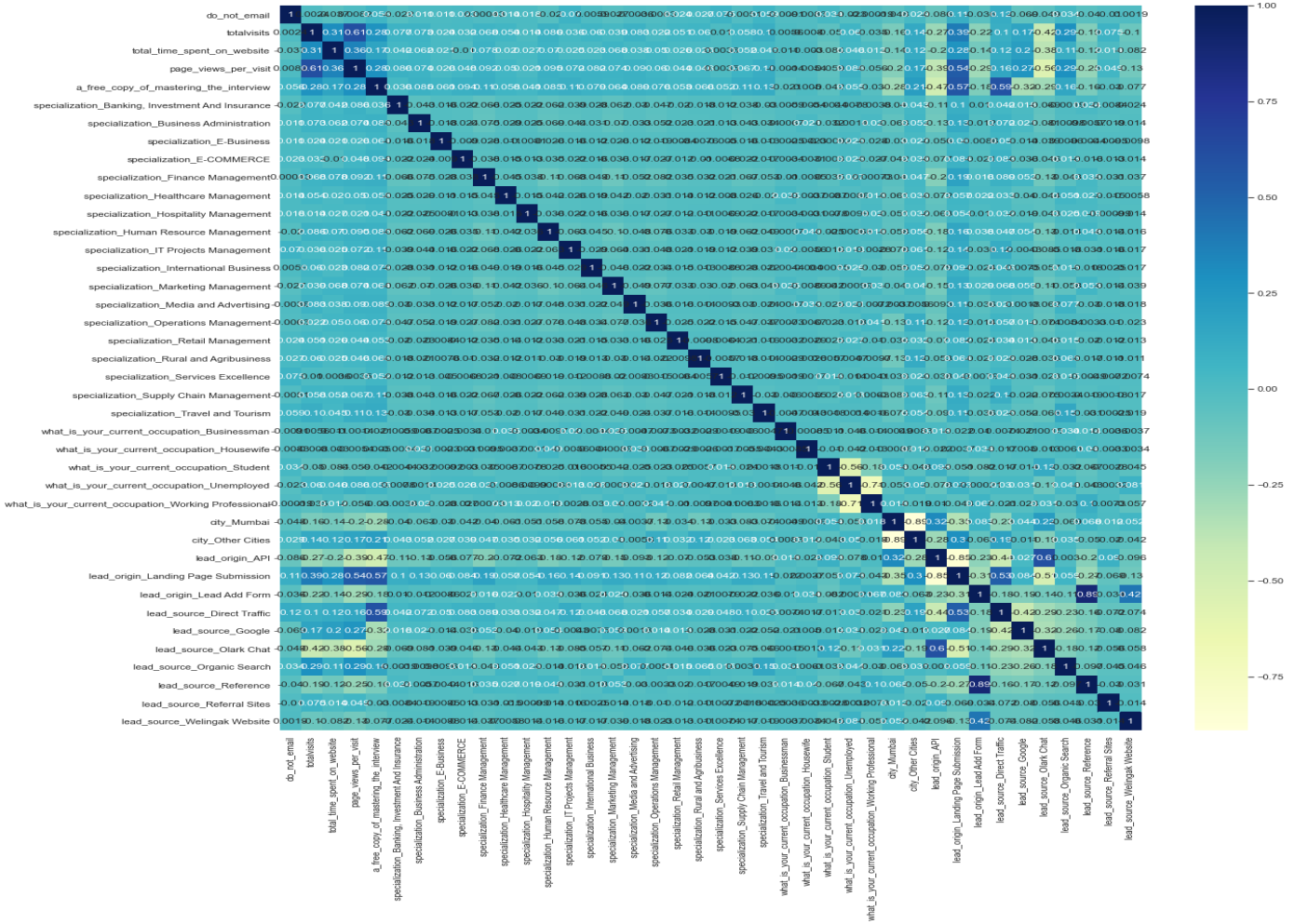
```
India                2918
gulf_country           32
United States          30
other_country          20
european_country       11
Name: country, dtype: int64
```

```
df1.country[df1.city=='Other Cities of Maharashtra'].value_counts()
```

```
India                391
gulf_country           7
other_country          6
United States          5
european_country       5
Name: country, dtype: int64
```

MACHINE LEANRING MODEL MAKING

After data cleaning, imputing, creating dummy variables and so on we started our model building process with nearly 40 variables. Which was nearly non-readable. There we proceed with RFE directly to ease the process of feature selection.



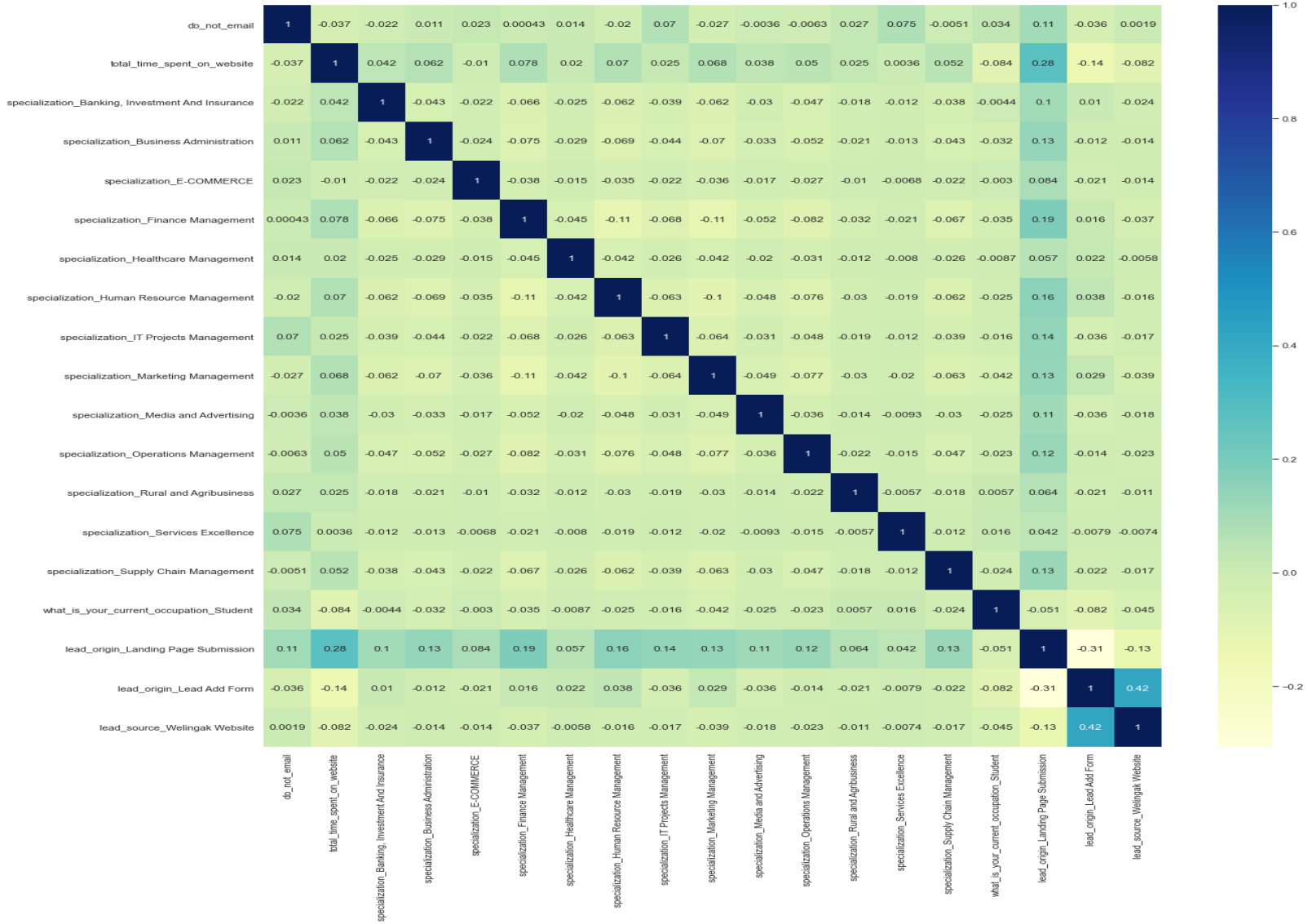
MACHINE LEANRING MODEL MAKING

RFE feature selection with 20 cut-off got us 20 most relevant variables with lowest p-value and lowest VIF. VIF was used to study the multi-collinearity of the feature variables.

'do_not_email',	'specialization_Marketing Management',
'total_time_spent_on_website',	'specialization_Media and Advertising',
'specialization_Banking, Investment And Insurance',	'specialization_Operations Management',
'specialization_Business Administration',	'specialization_Rural and Agribusiness',
'specialization_E-COMMERCE',	'specialization_Services Excellence',
'specialization_Finance Management',	'specialization_Supply Chain Management',
'specialization_Healthcare Management',	'what_is_your_current_occupation_Housewife',
'specialization_Human Resource Management',	'what_is_your_current_occupation_Student',
'specialization_IT Projects Management',	'lead_origin_Landing Page Submission',
'lead_origin_Lead Add Form',	'lead_source_Welingak Website'

MACHINE LEANRING MODEL MAKING

Heatmap reflecting the co-relation between each of the variables.



FINAL MACHINE LEARNING MODEL FOR X education SALES TEAM

With a probability threshold of 0.30 we were able to achieve the following score which when tested on the test data set matched well with the train data set. Indicating that it is a dependable model.

TRAIN DATA

Confusion matrics:

2982	983
554	1916

Accuracy score: **76%**

Sensitivity score: **77.57%**

Specificity score: **75.20%**

False Positive Rate: **24.79%**

TEST DATA

Confusion matrics:

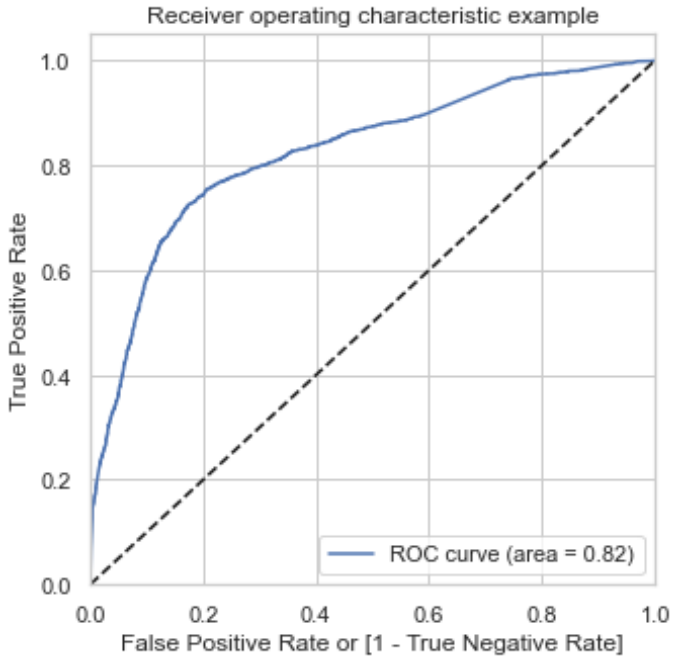
1261	437
233	828

Accuracy score: **75.71%**

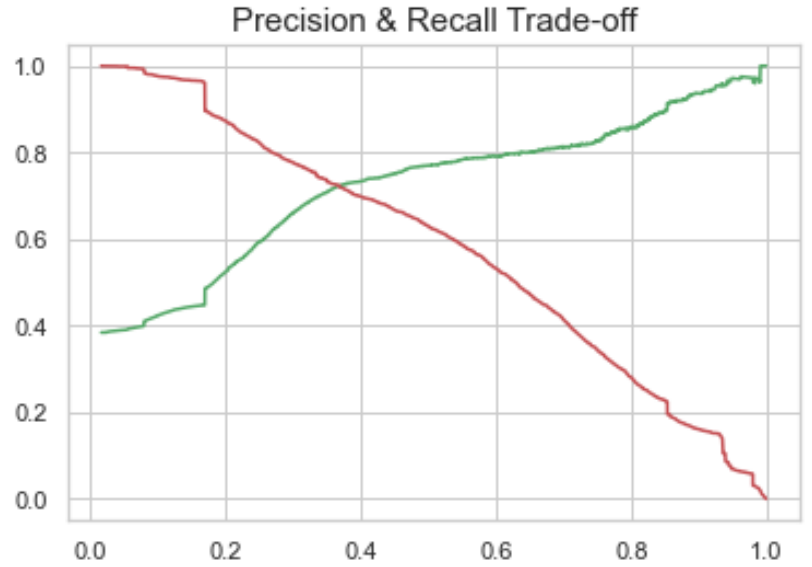
Sensitivity score: **78.03%**

Specificity score: **74.26%**

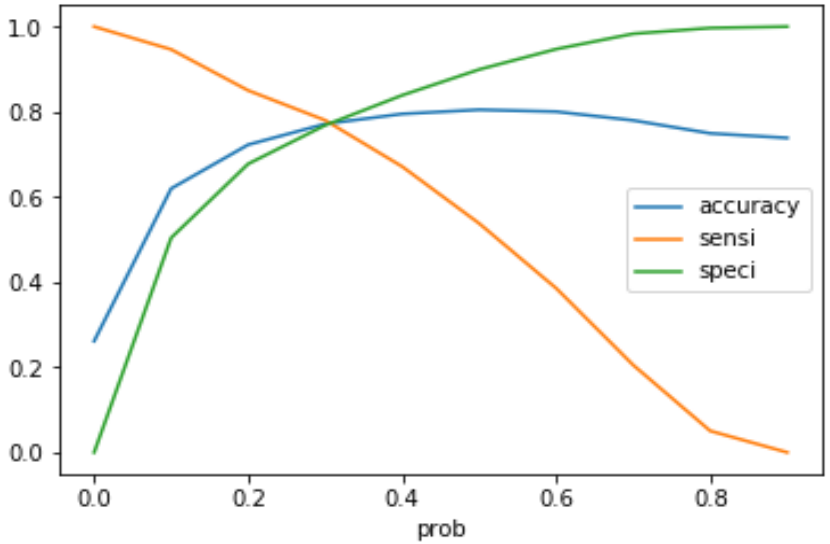
FINAL MACHINE LEARNING MODEL FOR X education SALES TEAM



ROC CURVE



OPTIMAL CUT-OFF POINT



SUGGESTION

GENERAL SUGGESTION

1. The initial data collection form needs some major rework
2. More options on providing various free reading or self preparing materials should be provided to gain the customer attention and interest.
3. Even though customers select 'NO' for call & e-mail they are highly convertible.

SPECIALIZATION BASED SUGGESTION

1. Banking, Investment And Insurance 1.1235
2. Business Administration 0.8662
3. E-Commerce 0.8385
4. Finance Management 0.8982
5. Healthcare Management 1.1608
6. Human Resource Management 0.9038
7. IT Projects Management 0.8173
8. Marketing Management 0.9700
9. Media and Advertising 0.8955
10. Operations Management 1.0477
11. Rural and Agribusiness 0.8370
12. Services Excellence 1.2927
13. Supply Chain Management 0.9176