Lead Score Case Study

Darpan Shah
VenkataNagarjuna Hebbar

Problem Statement

- An education company named X Education 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.
- 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%.

Methodology

Data Cleaning and Preparation

- Identifying the Quality of data and cleaning the data whenever is required
- Handling the “select” values and missing values
- Treating the Outliers

Building Model

- Splitting the data into train and Test data sets using 70-30 ratio
- Apply Logistic Regression using GLM model on train data with RFE
- Remove the features which are having P-value is greater than 0.05 and VIF 5

Inferences and Recommendations

- Based on the model, identifying the variables which can influence the objective
- Draw the recommendations based on the model

Data Cleaning

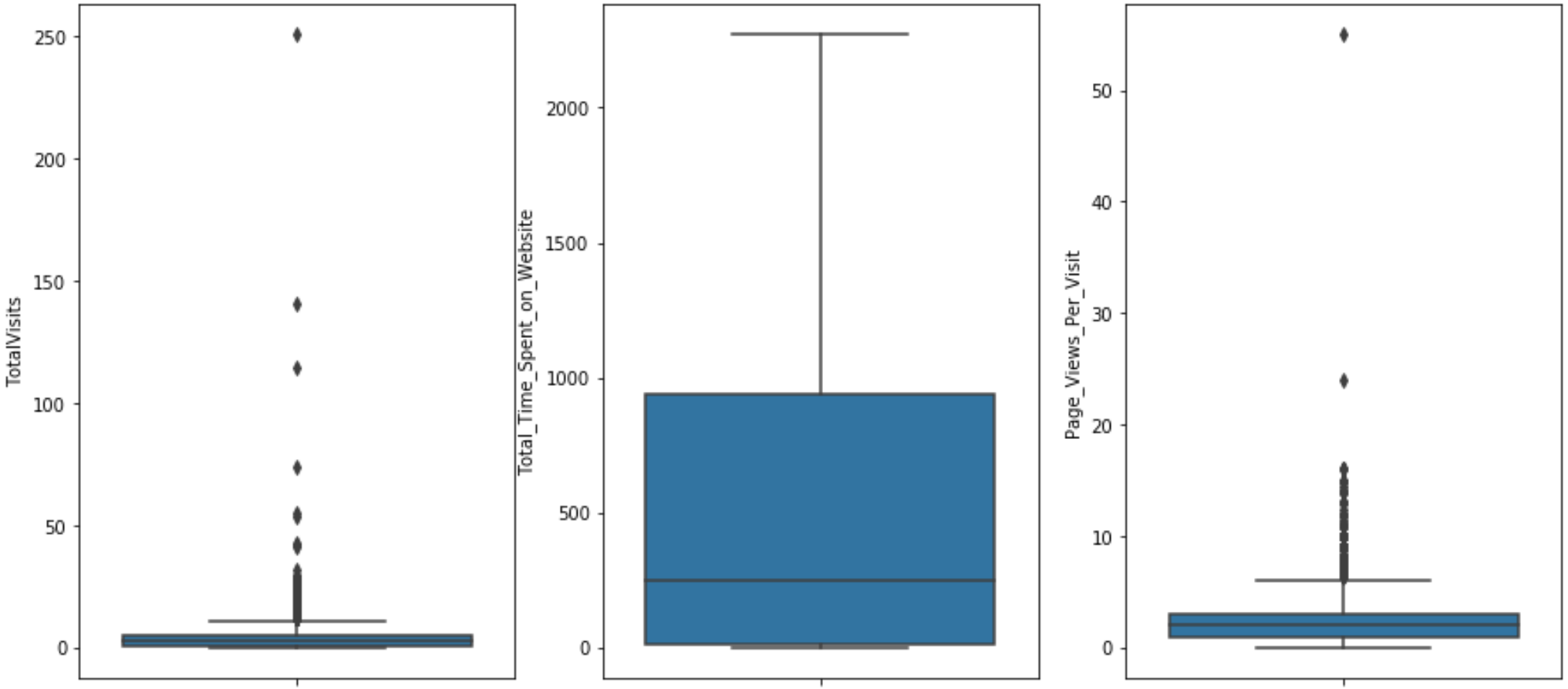
- Purged the columns where missing values are more than 30%
- Handling the “Select and NaN” values: There are few columns contains Select as one value which means customer not selected any value in that list, if we consider this as null value then we might be loosing almost 45% of data. Hence we are changing to “other” or “unknown” for the columns which are having Select value
- Purged columns which are having unique value as they don’t make much variance in model building.
- We checked the outlier data using Percentiles, Pairplot and Boxplots and treated them accordingly.

Missing Values

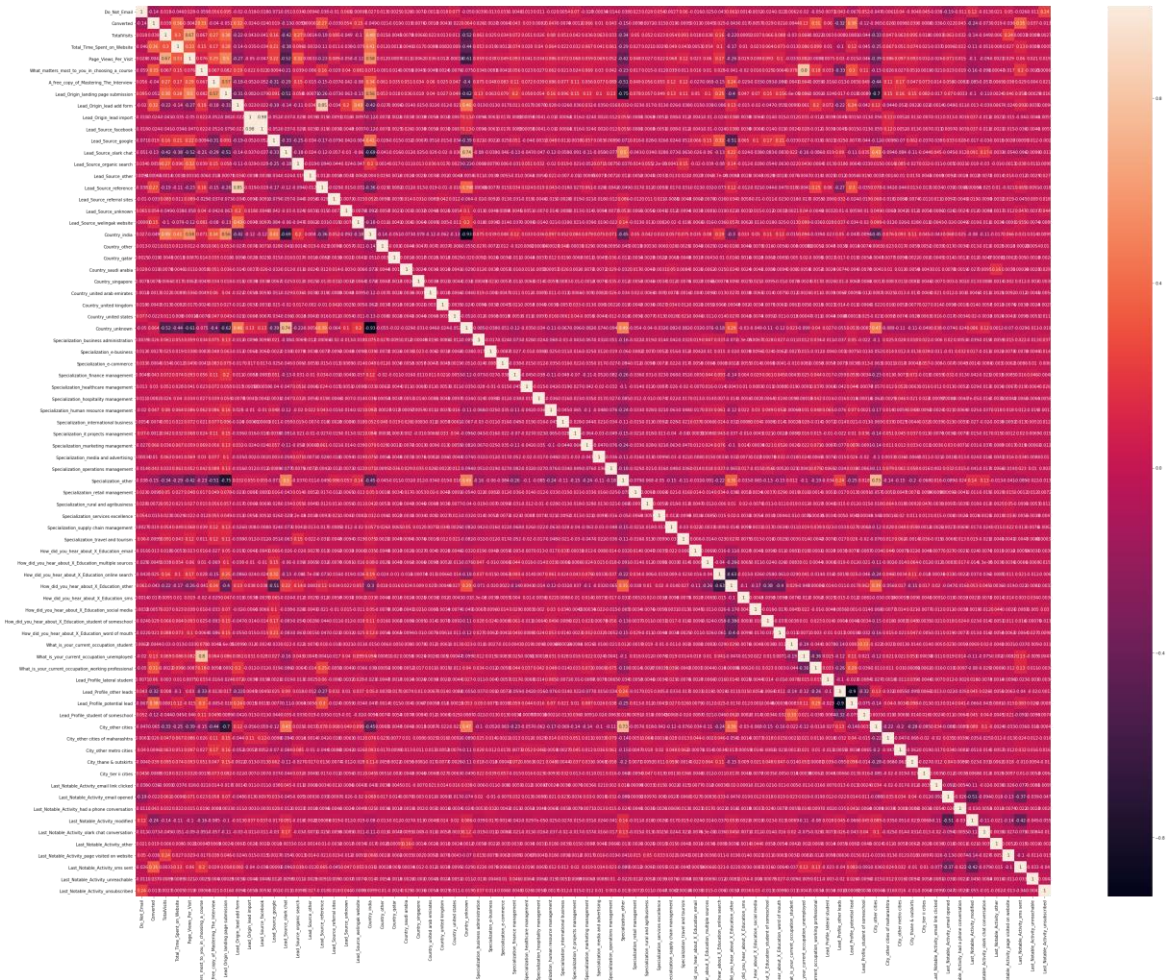
```
# checking the percentage of null values of each column.
round(100*(lead_score_df.isnull().sum()/len(lead_score_d
```

Lead_Quality	51.59
Asymmetrique_Profile_Score	45.65
Asymmetrique_Activity_Score	45.65
Asymmetrique_Profile_Index	45.65
Asymmetrique_Activity_Index	45.65
Tags	36.29
What_matters_most_to_you_in_choosing_a_course	29.32
Lead_Profile	29.32
What_is_your_current_occupation	29.11
Country	26.63
How_did_you_hear_about_X_Education	23.89
Specialization	15.56
City	15.37
TotalVisits	1.48
Page_Views_Per_Visit	1.48
Last_Activity	1.11
Lead_Source	0.39
Do_Not_Email	0.00
Do_Not_Call	0.00
Converted	0.00
Total_Time_Spent_on_Website	0.00
Lead_Origin	0.00
Lead_Number	0.00
Last_Notable_Activity	0.00
Newspaper_Article	0.00

Outlier Treatment



Variance between variables

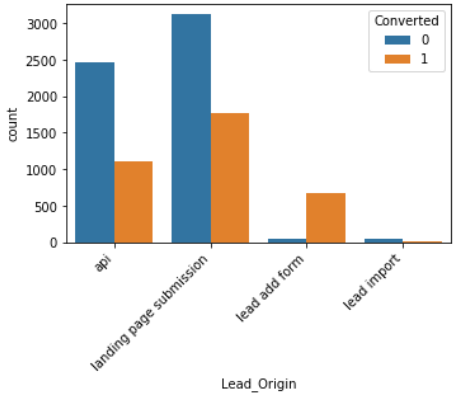


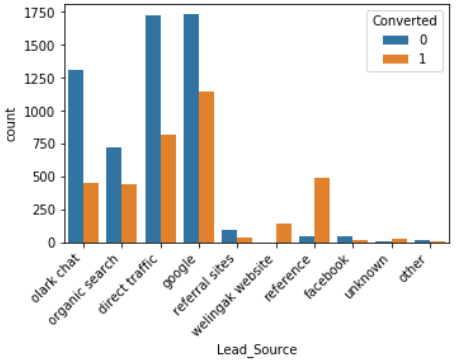
	Var 1	Var 2	coeff
3	Lead_Origin_lead import	Lead_Source_facebook	0.981709
5	Lead_Origin_lead add form	Lead_Source_reference	0.852594
6	Country_unknown	Lead_Source_olark chat	0.741415
8	What_is_your_current_occupation_unemployed	What_matters_most_to_you_in_choosing_a_course	0.798003
9	City_other cities	Specialization_other	0.728733

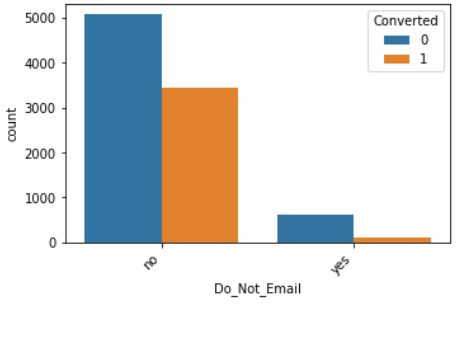
Model Building

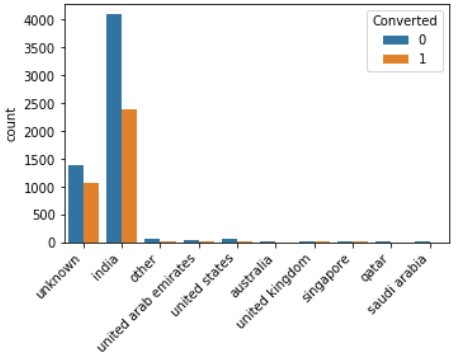
- Split data into train and test with 70-30 ratio and used RFE method to select the features
- Used StandardScaler to scale the features
- Run GLM Model on train data and verify the p-values and Variance Inflation Factor.
- find Confusion matrix, Accuracy Score.

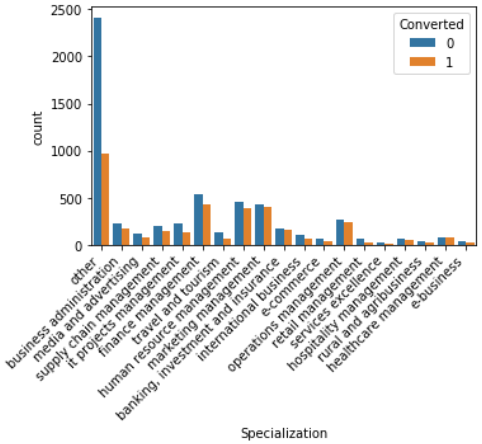
EDA

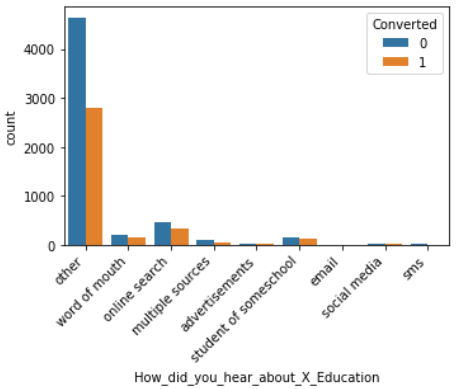

Converted
0
1
count
api
landing page submission
lead add form
lead import
Lead_Origin

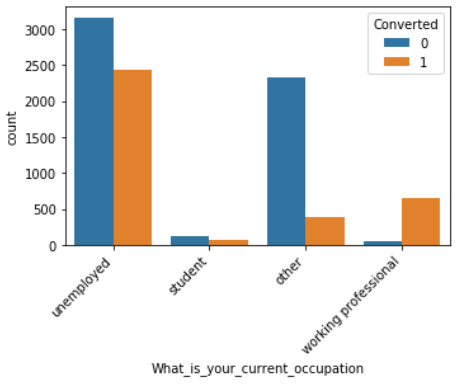

Converted
0
1
count
olark chat
organic search
direct traffic
google
referral sites
welingak website
reference
facebook
unknown
other
Lead_Source

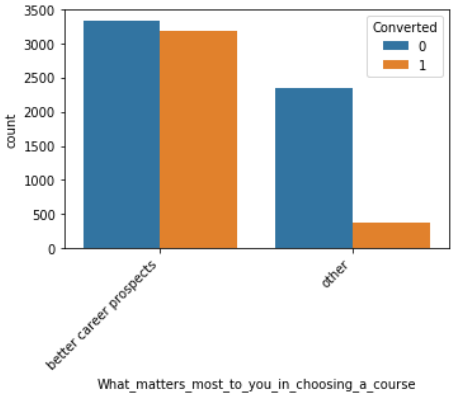

Converted
0
1
count
no
yes
Do_Not_Email

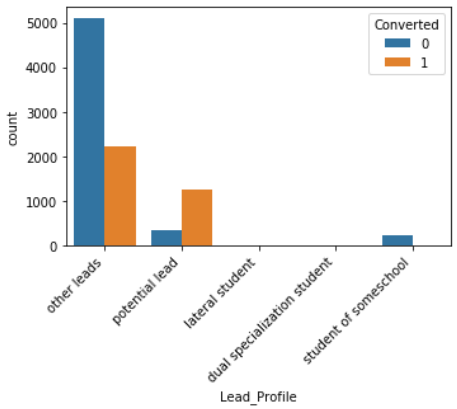

Converted
0
1
count
unknown
india
other
united arab emirates
united states
australia
united kingdom
singapore
qatar
saudi arabia
Country

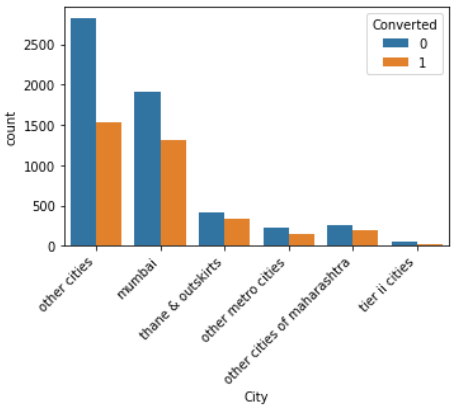

Converted
0
1
count
other
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization

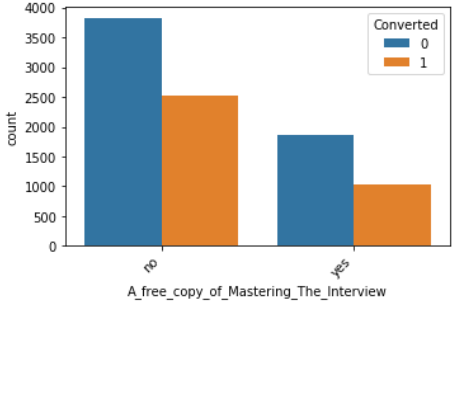

Converted
0
1
count
other
word of mouth
online search
multiple sources
advertisements
student of someschool
email
social media
sms
How_did_you_hear_about_X_Education

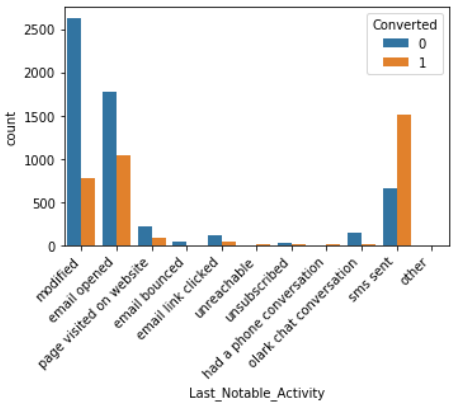

Converted
0
1
count
unemployed
student
other
working professional
What_is_your_current_occupation


Converted
0
1
count
better career prospects
other
What_matters_most_to_you_in_choosing_a_course


Converted
0
1
count
other leads
potential lead
lateral student
dual specialization student
student of someschool
Lead_Profile


Converted
0
1
count
other cities
mumbai
thane & outskirts
other metro cities
other cities of maharashtra
tier ii cities
City


Converted
0
1
count
no
yes
A_free_copy_of_Mastering_The_Interview


Converted
0
1
count
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
other
Last_Notable_Activity

Model Building

Generalized Linear Model Regression Results

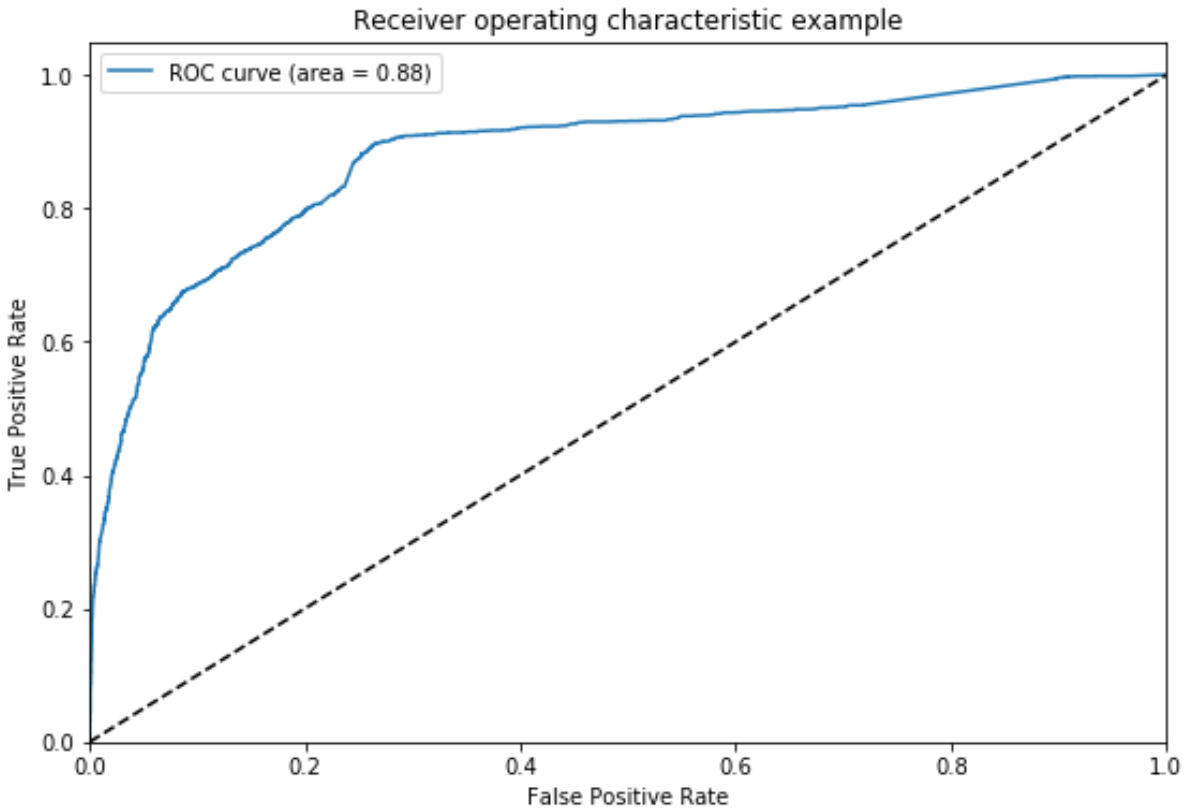
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6456
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2614.4
Date:	Mon, 02 Mar 2020	**Deviance:**	5228.7
Time:	19:42:13	**Pearson chi2:**	7.58e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

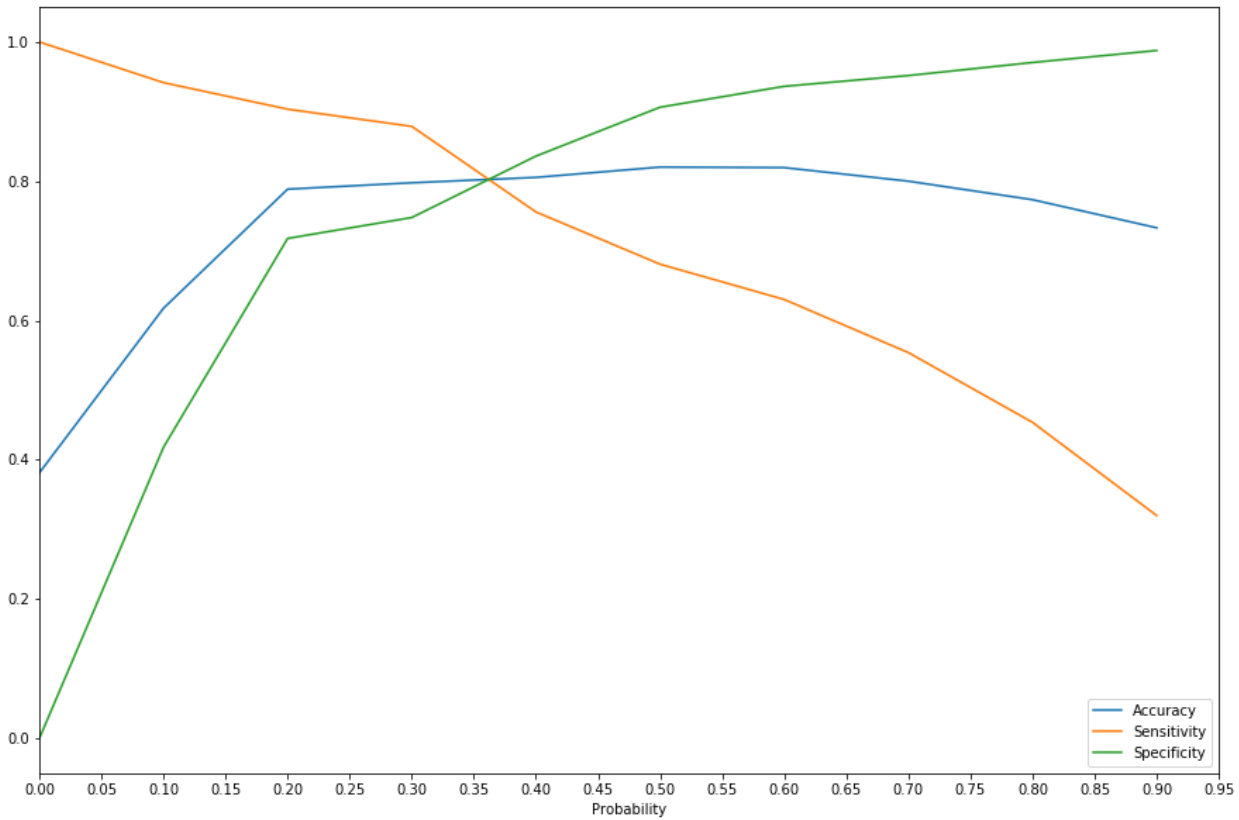
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1574	0.088	1.781	0.075	-0.016	0.331
Do_Not_Email	-1.4043	0.174	-8.081	0.000	-1.745	-1.064
Total_Time_Spent_on_Website	0.9175	0.035	26.104	0.000	0.849	0.986
Lead_Origin_lead add form	2.9471	0.190	15.474	0.000	2.574	3.320
Lead_Source_welingak website	2.6447	0.748	3.536	0.000	1.179	4.111
What_is_your_current_occupation_working professional	2.4252	0.192	12.664	0.000	2.050	2.800
Lead_Profile_other leads	-1.7475	0.095	-18.450	0.000	-1.933	-1.562
Lead_Profile_student of someschool	-3.3274	0.428	-7.771	0.000	-4.167	-2.488
Last_Notable_Activity_had a phone conversation	2.8195	1.150	2.451	0.014	0.565	5.074
Last_Notable_Activity_sms sent	1.6633	0.079	21.075	0.000	1.509	1.818
Last_Notable_Activity_unreachable	1.5023	0.546	2.751	0.006	0.432	2.573
Last_Notable_Activity_unsubscribed	1.3267	0.539	2.459	0.014	0.269	2.384

	Features	VIF
2	Lead_Origin_lead add form	1.38
5	Lead_Profile_other leads	1.32
8	Last_Notable_Activity_sms sent	1.32
3	Lead_Source_welingak website	1.24
0	Do_Not_Email	1.16
4	What_is_your_current_occupation_working profes...	1.15
10	Last_Notable_Activity_unsubscribed	1.07
1	Total_Time_Spent_on_Website	1.06
6	Lead_Profile_student of someschool	1.01
7	Last_Notable_Activity_had a phone conversation	1.00
9	Last_Notable_Activity_unreachable	1.00

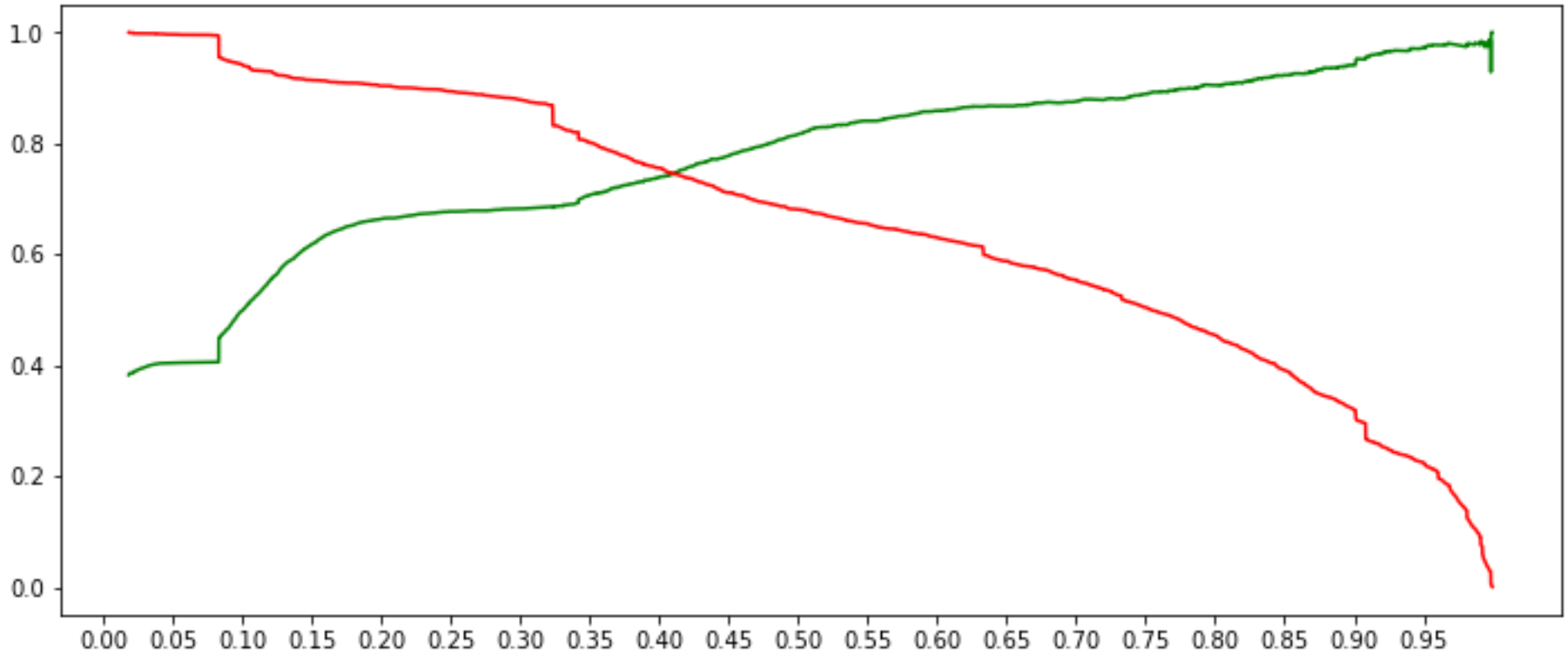
From the above, using RFE we have taken the variables and we removed the features manually where p value is more than 0.05 and VIF is greater than 5. we obtained the final model as above.

ROC Curve


Receiver operating characteristic example
ROC curve (area = 0.88)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy
Sensitivity
Specificity
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95

Precision Recall Tradeoff



As per the above, precision recall tradeoff is 0.41(after 0.40) hence we will consider the 0.45 as precision recall trade off

Train – Test Data results

Train - Accuracy , Precision and Recall

```
#Accuracy
metrics.accuracy_score(Y_train_pred_final.Conver
```

0.8123067408781695

```
# Precision
TP / float(TP + FP)
```

0.7774822695035462

```
# Recall
TP / float(TP + FN)
```

0.7112733171127331

```
#Specificity
TN / float(TN+FP)
```

0.8745627186406797

Test - Accuracy , Precision, Recall and Specificity

```
# Accuracy.
metrics.accuracy_score(Y_pred_final.Converted, Y_pred_final.f
```

0.8134920634920635

```
# Precision
TP / float(TP+FP)
```

0.7991718426501035

```
# Recall
TP / float(TP+FN)
```

0.7050228310502283

```
#Specificity
TN / float(TN+FP)
```

0.8843172331544424

Recommendations

- Leads with 0.45 and above should be considered ad Hot leads and Sales Team should focus on calling these leads to achieve maximum conversion s as paying customers.
- The Top Three variables in our Model which contributes the most towards the conversion of leads
  - **Lead origin**
  - **Last Notable Activity**
  - **Lead Source**

- The Top Three categorical variables should be focused most inorder to increase the probability are
  - **Lead_Origin_lead add form** with a coefficient of 2.9471
  - **Last_Notable_Activity_had a phone conversation** with a coefficient of 2.8195
  - **Lead_Source_welingak website** with a coefficient of 2.6447

Conclusions

- From the Logistic Model we built, We can conclude that X Education is to identify the leads that are most likely to be a paying customers.
- Higher the probability score means that the lead is hot, means the chances of converting into paying customer is high when the score is high.
- Accuracy and precision is more than 80%, it also help to meet the target lead conversion ratio of 80%.
- The model can be adjusted in future like:
  - A period of 2 months every year during which X education hire some interns and during this page, they wish to make the lead conversion more aggressive and want almost all of the potential leads to be converted, hence to make this we can lower the cutoff further. Lets say we can reduce the cutoff to 0.30 to 0.35 where the number of leads is 3974(0.35 cut-off)where as with 0.45 we have number of leads is 3222.
  - Once the company reaches its target for the quarter before the deadline and during this time, the company wants the sales team to focus on the other new work as week. Also during this time company's aim is not to make any phone calls unless its extremely necessary. To achieve this, we need to minimize the phone calls during this time, we can raise the cut-off value higher than 0.45 to target less customer but with high conversion rate. We will set the cut-off value around 0.85 to 0.90 with projected leads as 1468 to 1180 (0.85 cutoff) , as the precision is high we are contacting the high chances of conversion and will achieve the aim to avoid phone calls.