Lead Score Assignment

By Kamesh Vishwakarma & A V Ramya Keerthana

Contents-

1. Problem Statement

An education company named X Education sells online courses to industry professionals. 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.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

What we need to do ?

1. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
2. The company requires you to build a model wherein you need to assign a lead score 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

2. Overall Approach

1. **Reading and Understanding Data**
    1. Importing libraries
    2. Checking shape, description and info of dataset

2. **Data Preparation**
    1. ‘Select’ replacement
    2. Null Check
    3. Null Treatment
3. **Exploratory Data Analysis(EDA)**
    1. Defining Custom functions for visualization
    2. Variable Analysis with respect to target variable
    3. Dropping unnecessary columns
    4. Converting Binary Variables ‘yes’/’no’ to 1/0
    5. Checking correlation between numerical variables
4. **Dummy Variable creating from category variables for model generation**

2. Overall Approach

5. Train Test Split

- 5.1 Importing library
- 5.2 Creating X and y data set.
- 5.3 Splitting data into Train and Test data set
- 5.4 Feature Scaling

6. Model Building

- 6.1 Feature selection using RFE
- 6.2 Assessing the model with Statsmodels
- 6.3 Predicting values with model built

7. Model Evaluation

- 7.1 Confusion matrix
- 7.2 Accuracy (Train set)
- 7.3 Metrics other than accuracy
- 7.4 ROC curve

2. Overall Approach

8. Optimal Cut-off Point

9. Assigning Lead Score to leads in Train set

10. Precision and Recall TradeOff

11. Prediction on Test set

12. Model Evaluation on Test set

13. Generating Hot leads

14. Conclusion

READING AND UNDERSTANDING DATA

1. Data set had total 9240 rows and 37 columns.
2. Out of which 6 columns were numeric rest were categorical.
3. Numerical columns description told that there are some outliers in numerical columns.

Numerical columns description

```
In [15]: df.describe([0.05, 0.25, 0.5, 0.9 , 0.95, 0.99])
```

Out[15]:

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
5%	582869.900000	0.000000	0.000000	0.000000	0.000000	12.000000	14.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
90%	650506.100000	1.000000	7.000000	1380.000000	5.000000	16.000000	19.000000
95%	655404.050000	1.000000	10.000000	1562.000000	6.000000	17.000000	20.000000
99%	659592.980000	1.000000	17.000000	1840.610000	9.000000	17.000000	20.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

2. Data Preparation

- 1. Replaced ‘select’ values with ‘np.nan’.
- 2. Null check
- 3. Null values treatment
  - Dropped columns with more than 40% null values.
  - Imputed null values with mode in categorical variables if null values are around 5-15%
  - Introduced new category as others if null values are around 20-30%

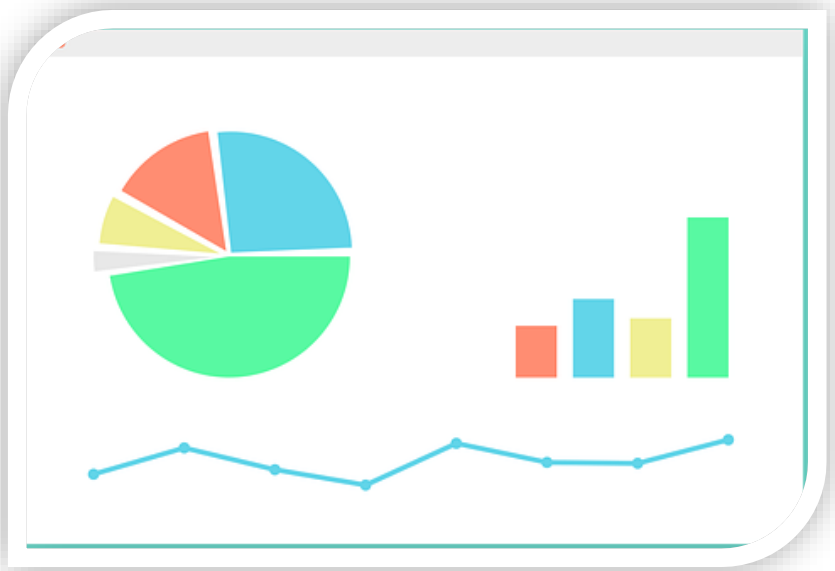
Null value percentage in various columns

In [19]:
```
# number's dont tell much let's do it in percentage and only columns with null values
# getting 'em in decreasing order might help too
df.isnull().mean()[(df.isnull().mean() * 100)>0].sort_values(ascending = False) * 100
```

Out[19]:

How did you hear about X Education	78.463203
Lead Profile	74.188312
Lead Quality	51.590909
Asymmetrique Profile Score	45.649351
Asymmetrique Profile Index	45.649351
Asymmetrique Activity Index	45.649351
Asymmetrique Activity Score	45.649351
City	39.707792
Specialization	36.580087
Tags	36.287879
What matters most to you in choosing a course	29.318182
What is your current occupation	29.112554
Country	26.634199
Page Views Per Visit	1.482684
TotalVisits	1.482684
Last Activity	1.114719
Lead Source	0.389610

dtype: float64



3. EDA – Exploratory Data Analysis

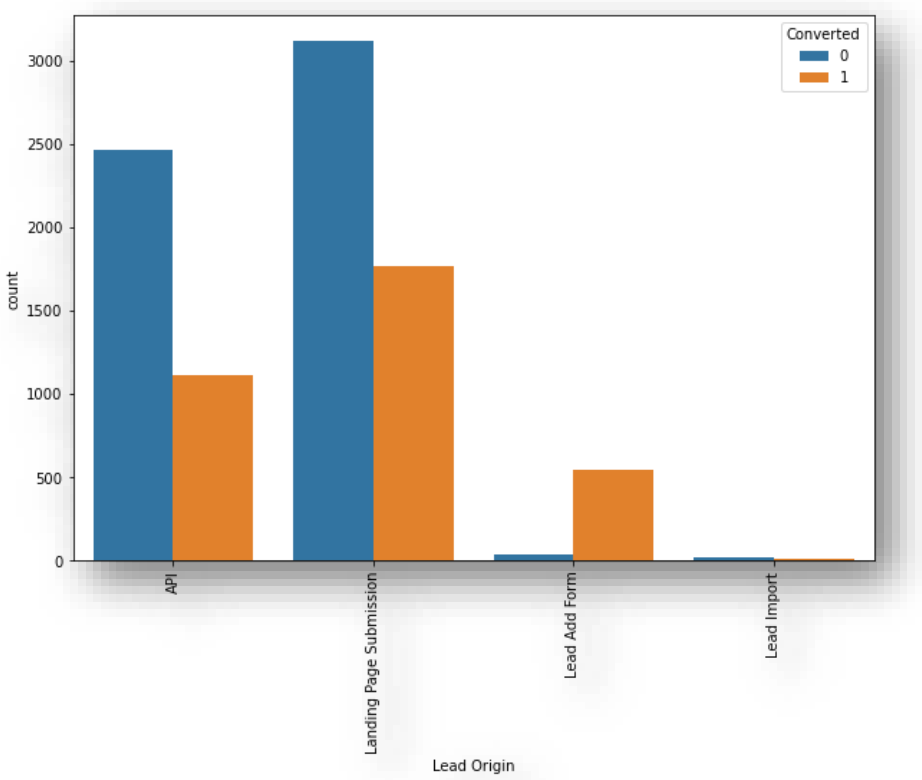
Variable analysis with respect to Target Variable

- **Lead Origin:**

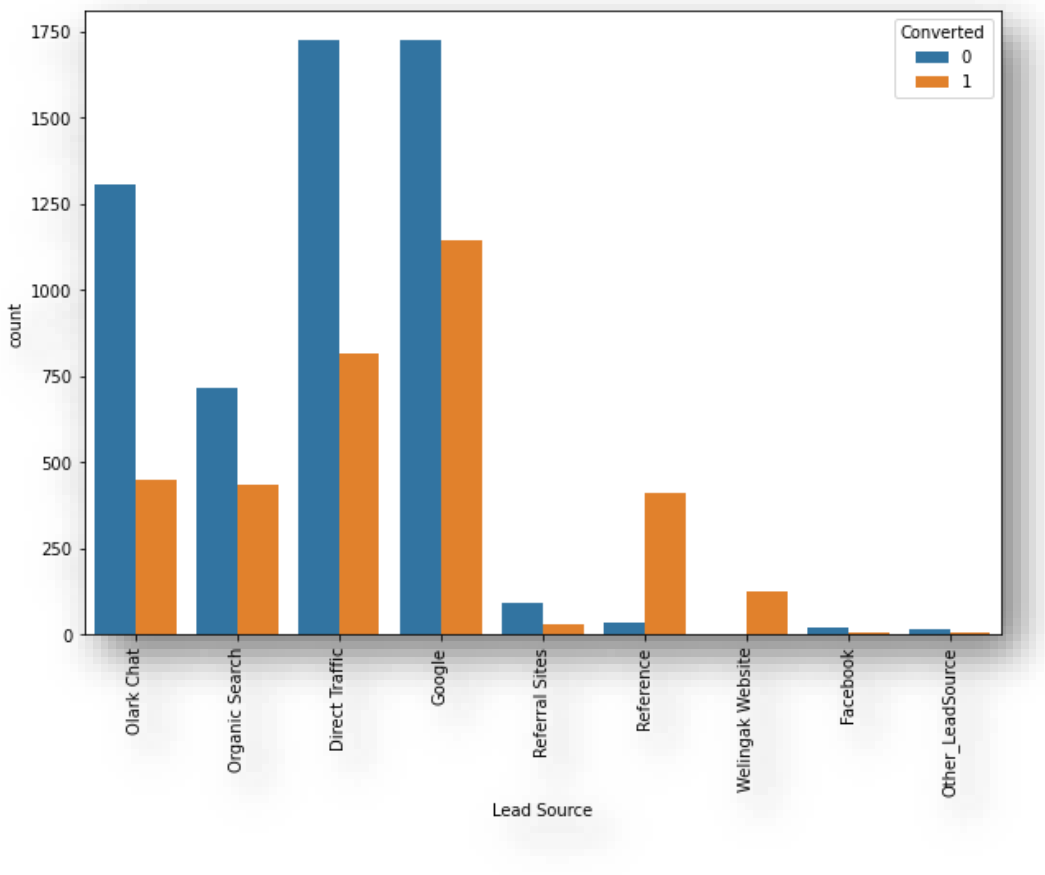
Inference:

- API and Landing Page Submission have nearly 30 - 35 % of lead conversion, but count is very high.
- Lead Add Form have nearly 90% conversion but count is low
- Lead Import do not have much leads and conversion, it is least of them all

Looks like we need to focus on API and Landing Page Submission to get more leads, and we need to increase count from Lead Add Form.



- **Lead Source:**



Inference:

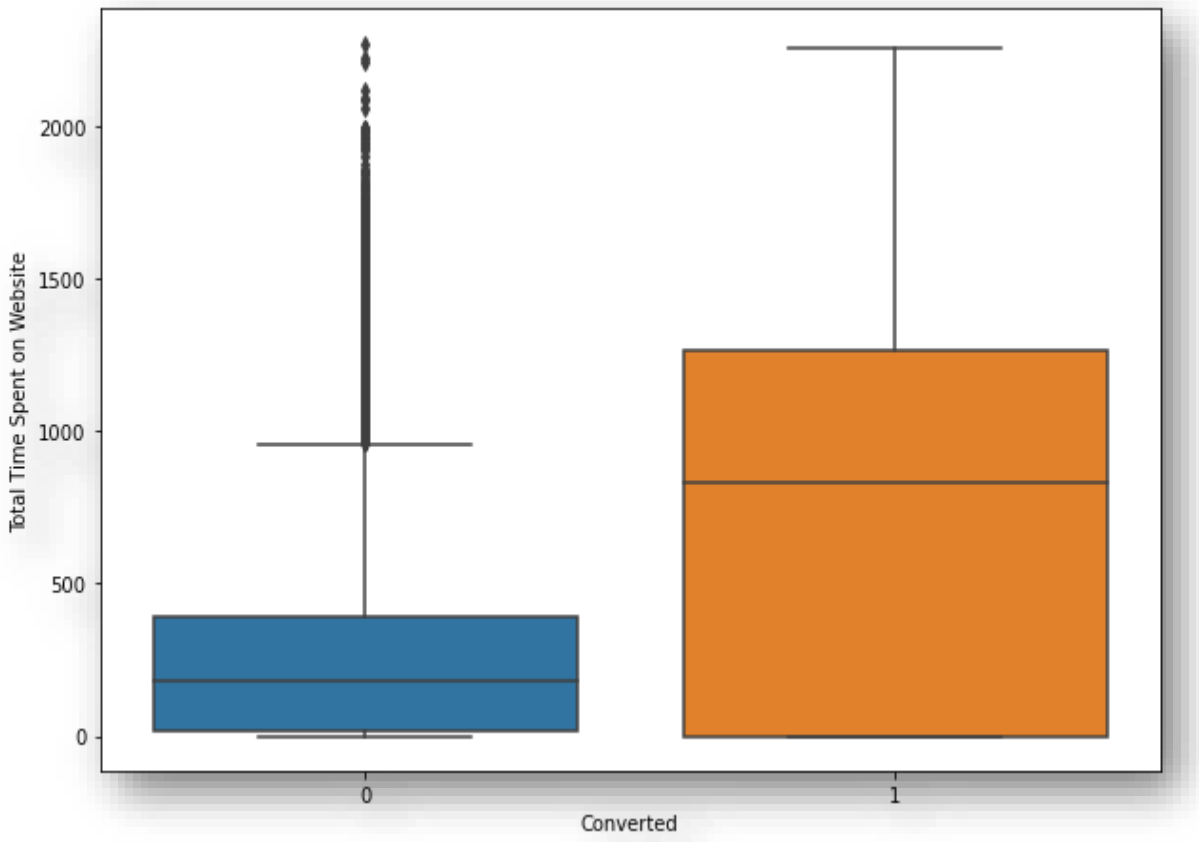
- API and Landing Page Submission have nearly 30 - 35 % of lead conversion, but count is very high.
- Lead Add Form have nearly 90% conversion but count is low
- Lead Import do not have much leads and conversion, it is least of them all

Looks like we need to focus on API and Landing Page Submission to get more leads, and we need to increase count from Lead Add Form.

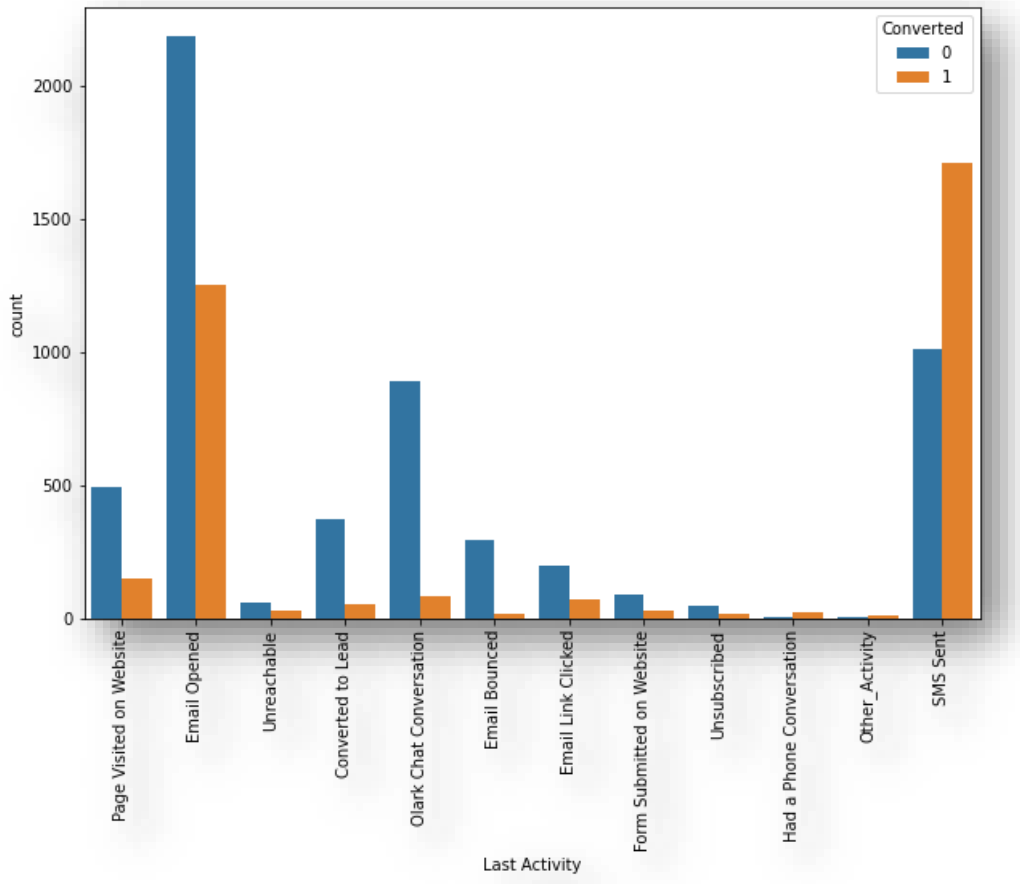
- **Total Time Spent on Website:**

Inference:

- So this tells that more time a lead spend on website more likely it is to convert. company should focus on their website UI/UX as well I think and try to make website more engaging.



Last Activity:



Inference:

- Most of the leads have opened their Emails, and nearly 30% of them converted as well.
- SMS sent leads have almost 60% lead conversion rate.

focusing on these two type of leads is important as it is critical point for leads to convert. It will work as a guidance when in progress leads. As which to focus more.

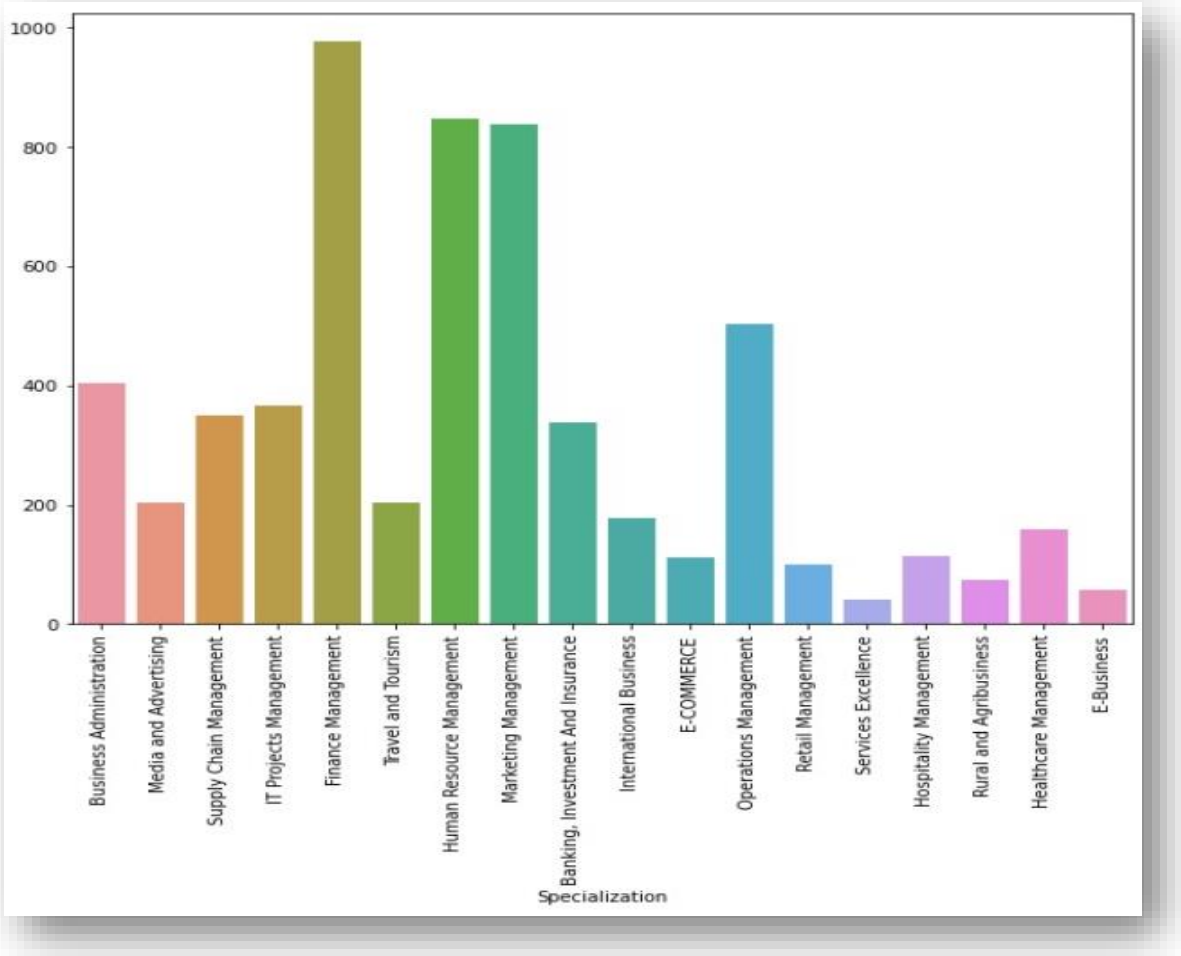
• Specialization :

Inference:

- Others are those leads who did not specify there specialization, their conversion rate is approximately 20%

- Should focus on leads with high conversion rate where specialization is specified.

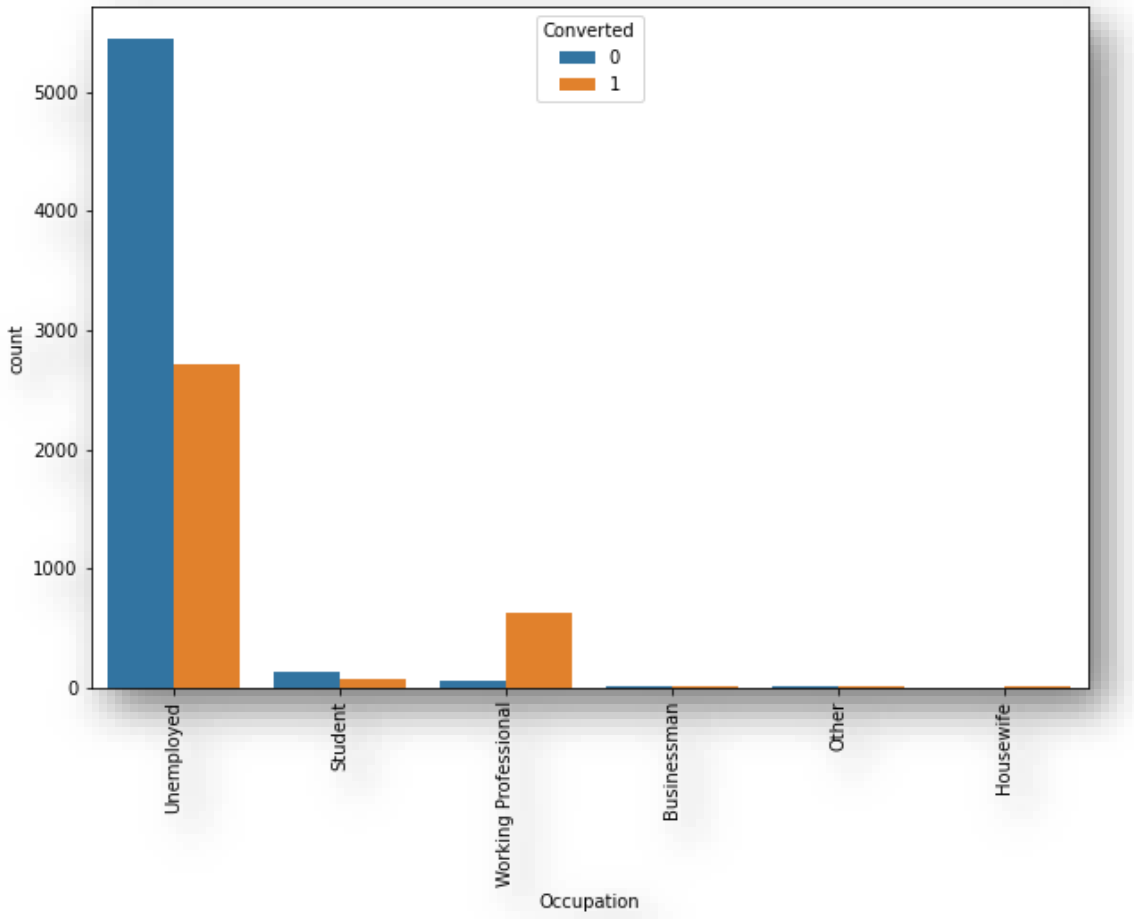
Finance Management, H R Management, Marketing Management, Operations management are specialization from where there is high chance for leads to convert. company should focus on these type of leads.



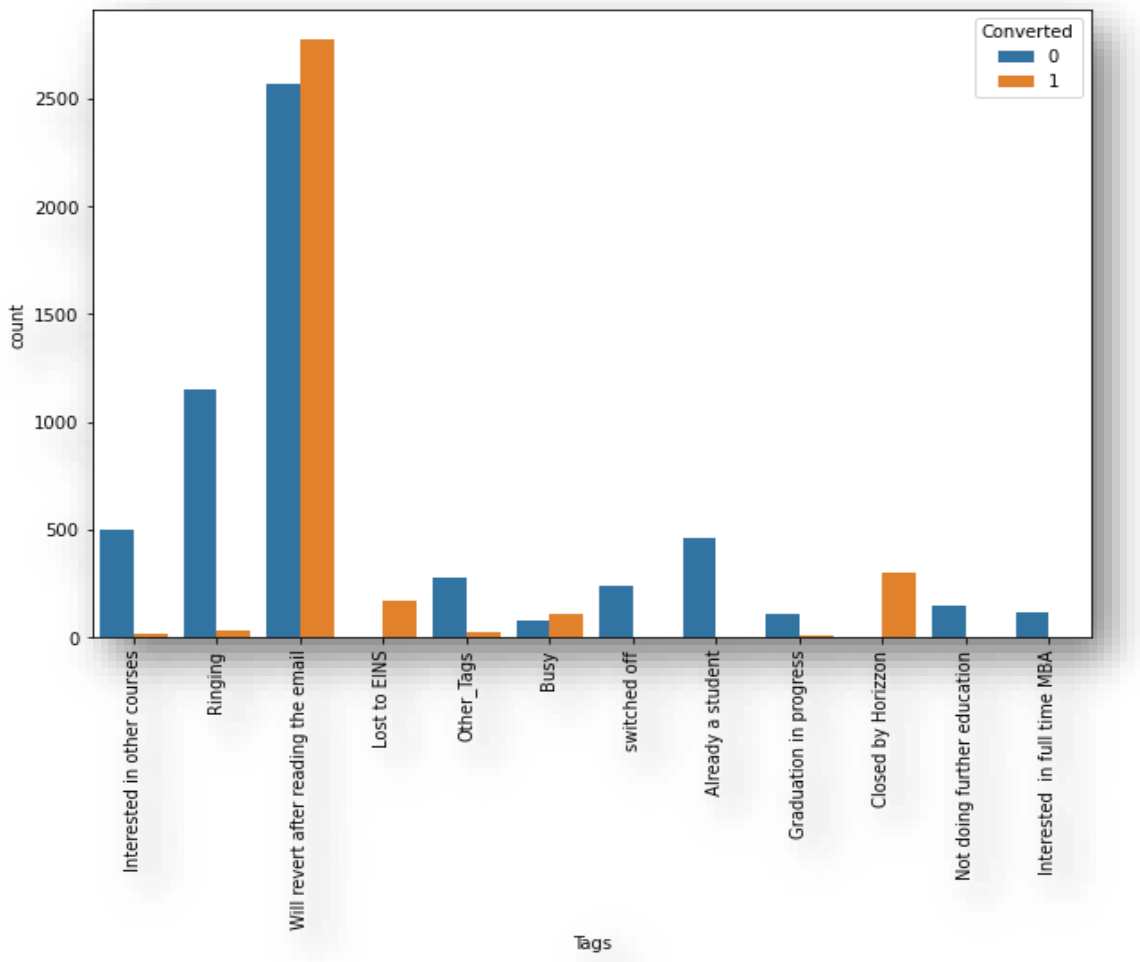
- **Occupation:**

Inference:

- Working professionals have highest chance of getting converted. Their conversion rate is more than 90%.
- Unemployed leads have most count but their conversion rate is among 30-35%.



- **Tags:**



inference:

- leads have more chances to convert if tag is 'will revert after reading email'. This type of leads have more than 50% of conversion rate.

- Lost to EINS, Busy, Closed by Horizon have high conversion rates but low count.

Focusing on leads with tag 'will revert after reading email' will result in more lead conversion. Should try to increase count of other high conversion tags.

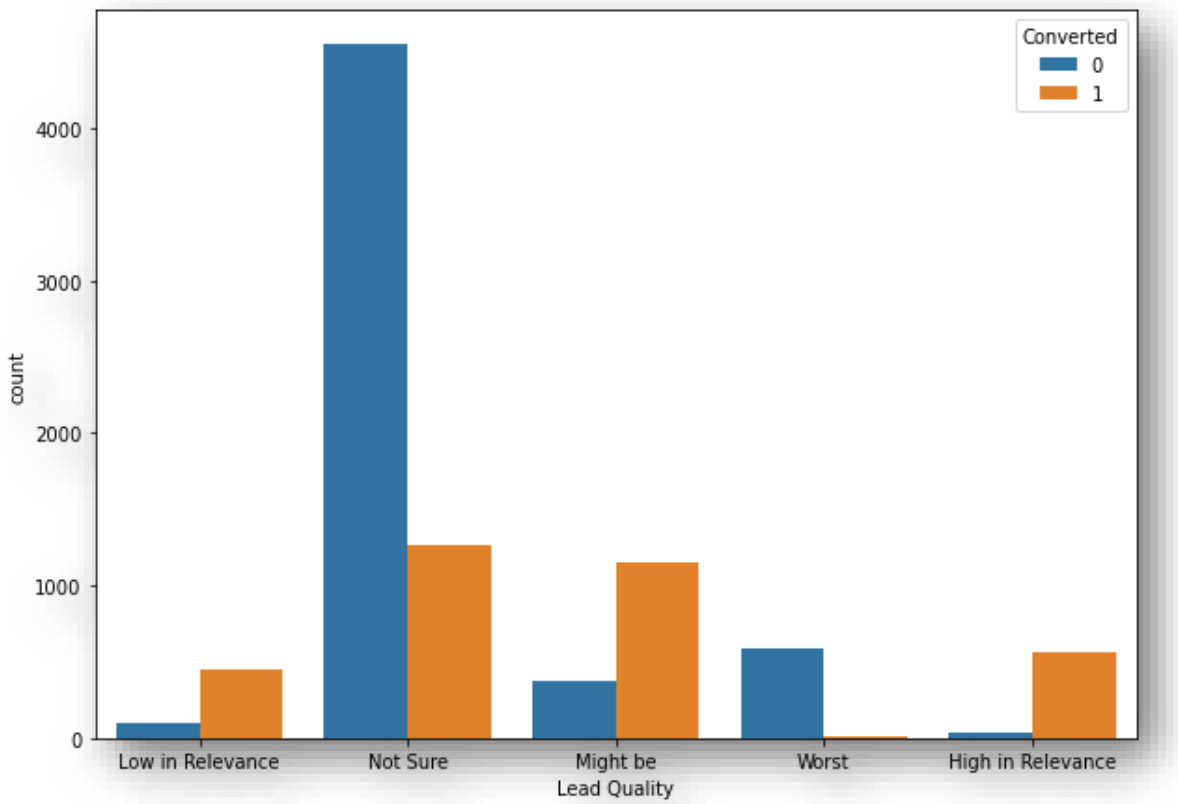
- **Lead Quality:**

inference:

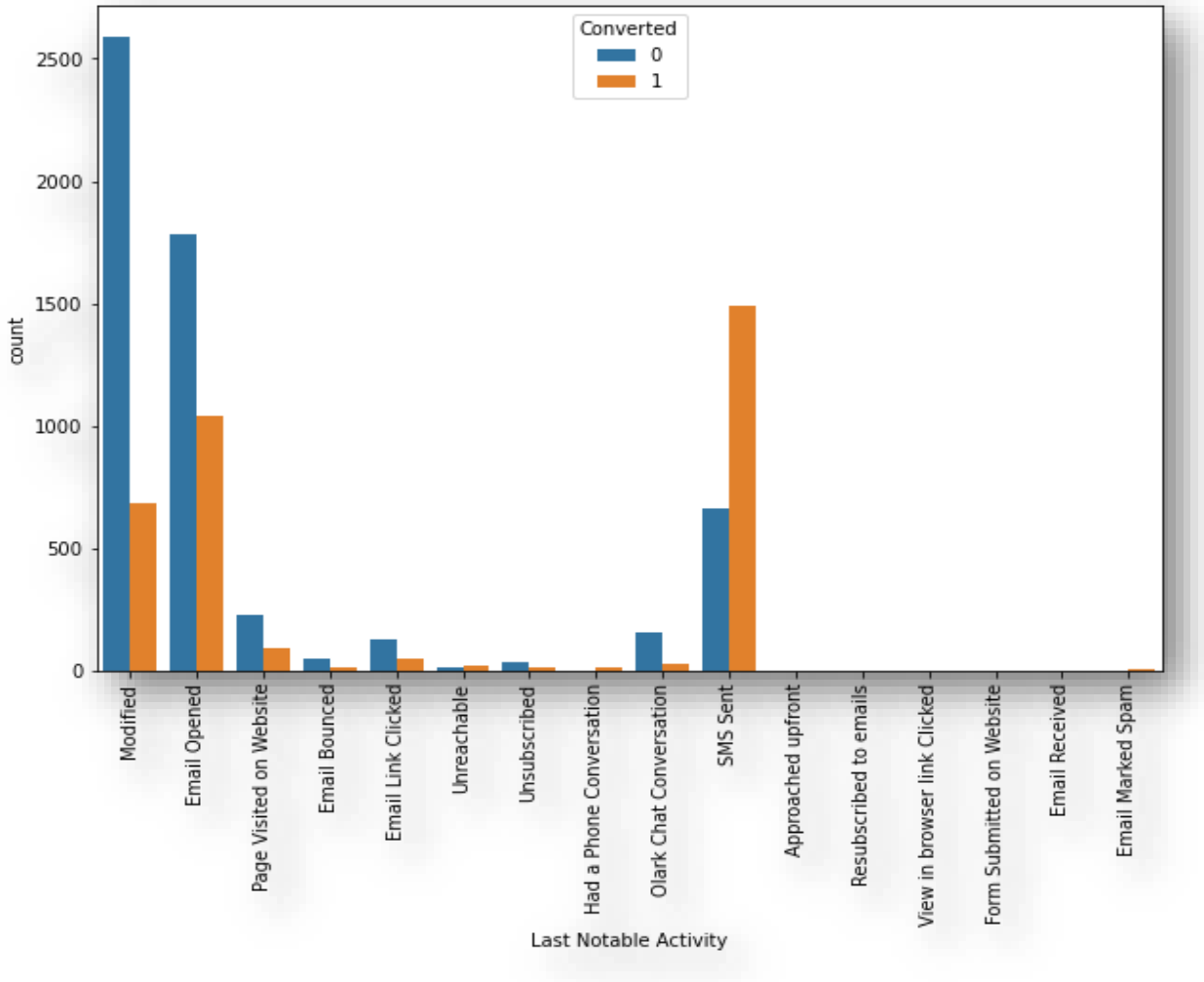
- As expected Not sure have high count but low conversion rate. and worst have lowest conversion rate.

- High in relevance have highest conversion rate, followed by Low in relevance and might be

when lead quality is in relevance, it is highly likely to convert leads.



- **Last Notable Activity:**



inference:

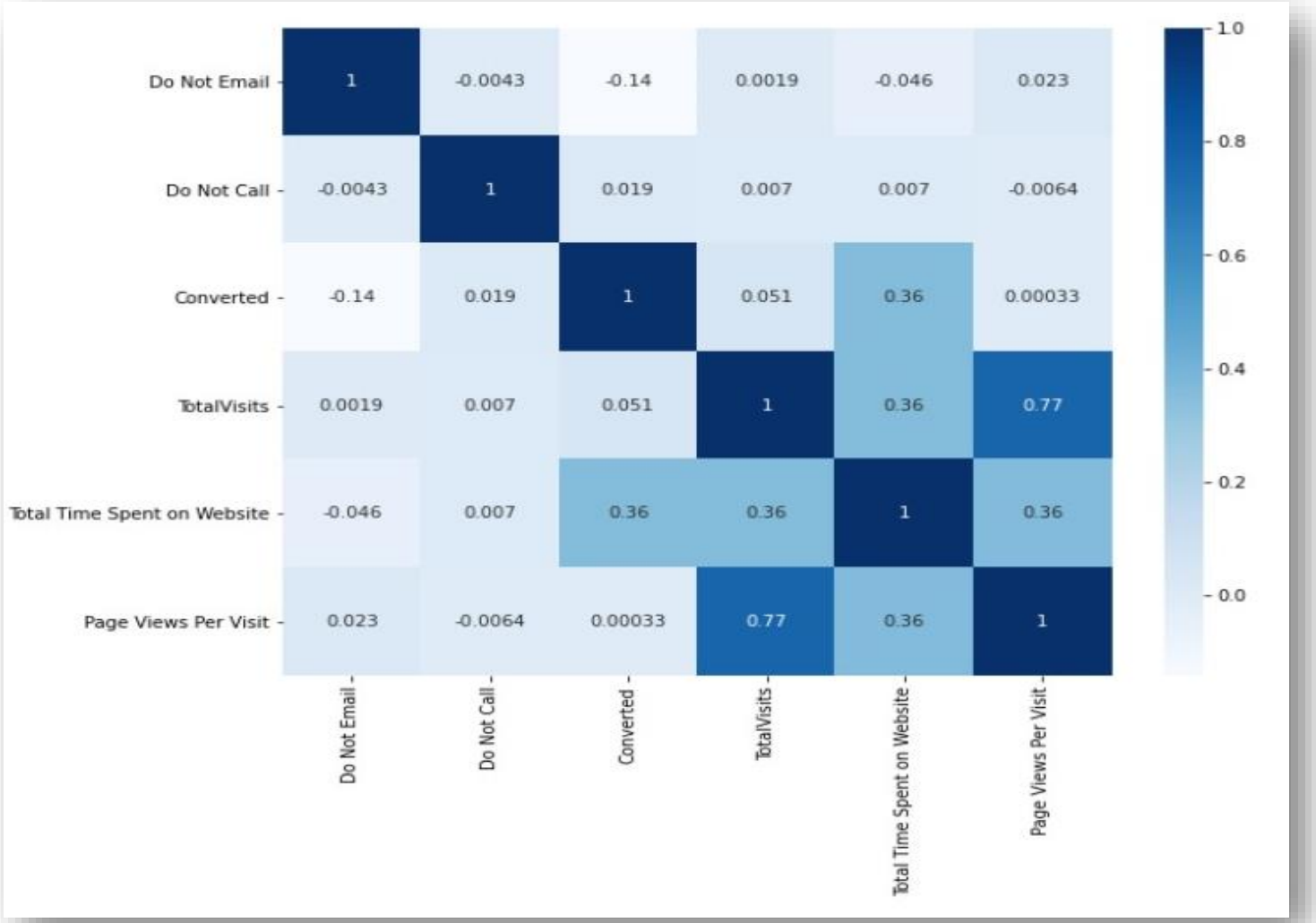
- leads where SMS is sent have highest chance of conversion, with conversion rate of around 70%.
- Most leads counts are from Modified and Email opened as last notable activity. If possible company should try to convert these leads.

Dropping Unnecessary columns

These columns were not useful in getting any useful inferences, so dropped.

'Lead Number', 'What matters most to you in choosing a course',
'Search', 'Magazine', 'Newspaper Article',
'X Education Forums', 'Newspaper', 'Digital Advertisement',
'Through Recommendations', 'Receive More Updates About Our Courses',
'Update me on Supply Chain Content', 'Get updates on DM Content',
'agree to pay the amount through cheque', 'A free copy of Mastering The Interview', 'Country'.

Correlation among numerical variables



5. Train – Test Split

- Created X and y data sets where X data set contains all independent variables, and y dataset contains Dependent variable.

- Split both data sets into 70:30 ratio where 70% data is for training and 30% data is for testing the model.

- Scaled numerical features.

6. Model Building

- Feature Selection through RFE method:

```
1 global col
2 col = X_train.columns[rfe.support_]
3 col
```

```
Index(['Do Not Email', 'Total Time Spent on Website',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Source_Olark Chat', 'Lead Source_Referral Sites',
       'Lead Source_Welingak Website', 'Specialization_Services Excellence',
       'Occupation_Housewife', 'Occupation_Unemployed',
       'Occupation_Working Professional', 'Lead Quality_Low in Relevance',
       'Lead Quality_Might be', 'Lead Quality_Not Sure', 'Lead Quality_Worst'],
      dtype='object')
```

Dropping sales variables-

The data we have consists of two categories –

1. Form generated data
2. Sales team generated.

This means that data is being updated after sales team start approaching leads. As we know our aim is to build a model to find leads that can convert into sales. So we have to neglect all the feature variables introduced by sales team.

those features are - [tags, Lead Quality, Last Activity, Last Notable Activity, Lead Profile, Asymmetric Activity Index, Asymmetric profile Index, Asymmetric Activity Score, Asymmetric Profile Score]

as we already have dropped some of the columns from data set.

Time to get rid of remaining as well.

[Tags, Lead Quality, Last Activity, Last Notable Activity]

Final Model:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2367.0
Date:	Mon, 17 May 2021	**Deviance:**	4734.1
Time:	20:03:02	**Pearson chi2:**	6.23e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.5861	0.247	10.465	0.000	2.102	3.070
Do Not Email	-1.2948	0.183	-7.083	0.000	-1.653	-0.936
Total Time Spent on Website	1.1178	0.043	25.855	0.000	1.033	1.203
Lead Origin_Landing Page Submission	-0.3566	0.094	-3.802	0.000	-0.540	-0.173
Lead Origin_Lead Add Form	2.7020	0.257	10.507	0.000	2.198	3.206
Lead Source_Olark Chat	0.9299	0.123	7.569	0.000	0.689	1.171
Lead Source_Referral Sites	-0.9376	0.371	-2.527	0.012	-1.665	-0.210
Lead Source_Welingak Website	3.6620	0.759	4.827	0.000	2.175	5.149
Occupation_Working Professional	1.8429	0.207	8.892	0.000	1.437	2.249
Lead Quality_Low in Relevance	-1.7063	0.279	-6.117	0.000	-2.253	-1.160
Lead Quality_Might be	-1.9261	0.250	-7.719	0.000	-2.415	-1.437
Lead Quality_Not Sure	-4.0446	0.241	-16.789	0.000	-4.517	-3.572
Lead Quality_Worst	-6.5935	0.439	-15.012	0.000	-7.454	-5.733

- Values are predicted using this model

7. Model Evaluation

Actual / Predicted	Not Converted	Converted
Not Converted	3334	571
Converted	523	1923

Confusion matrix

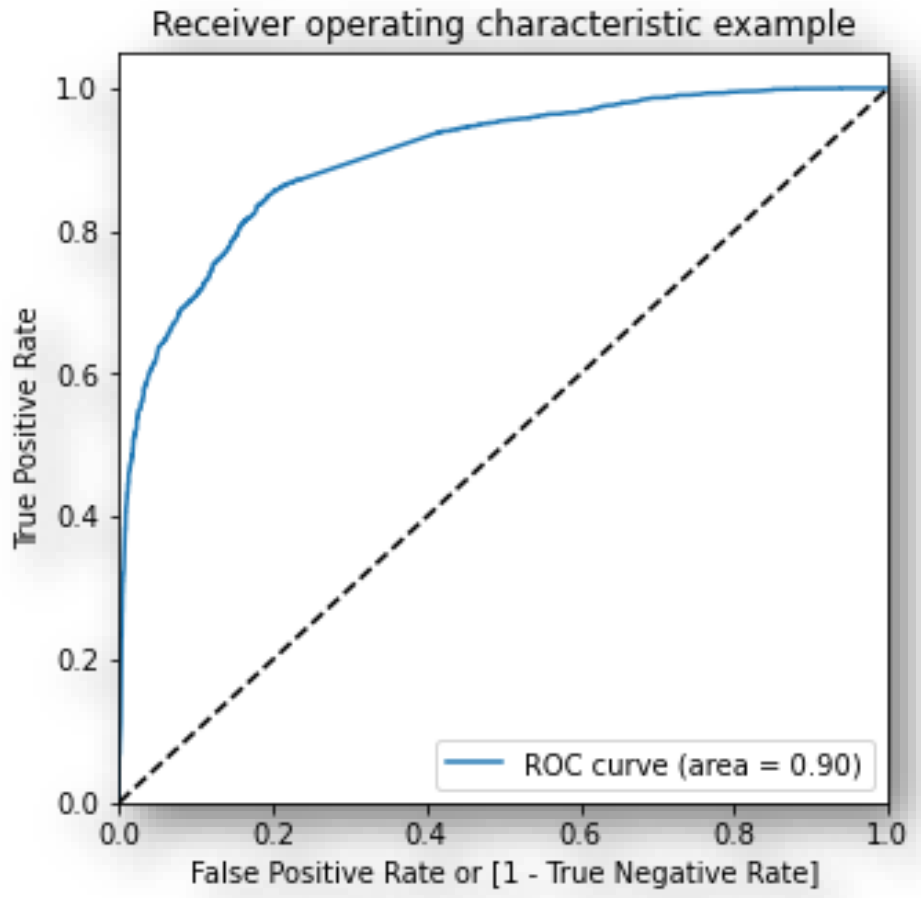
Other metrics on train set

with cutoff as 0.4

Metric	Value
Sensitivity	78.61%
Specificity	85.37%
False Positive Rate	14.62%
Positive predictive value	77.10%
Negative Predictive	86.44%
Acccuracy	82.07%

Plotting ROC

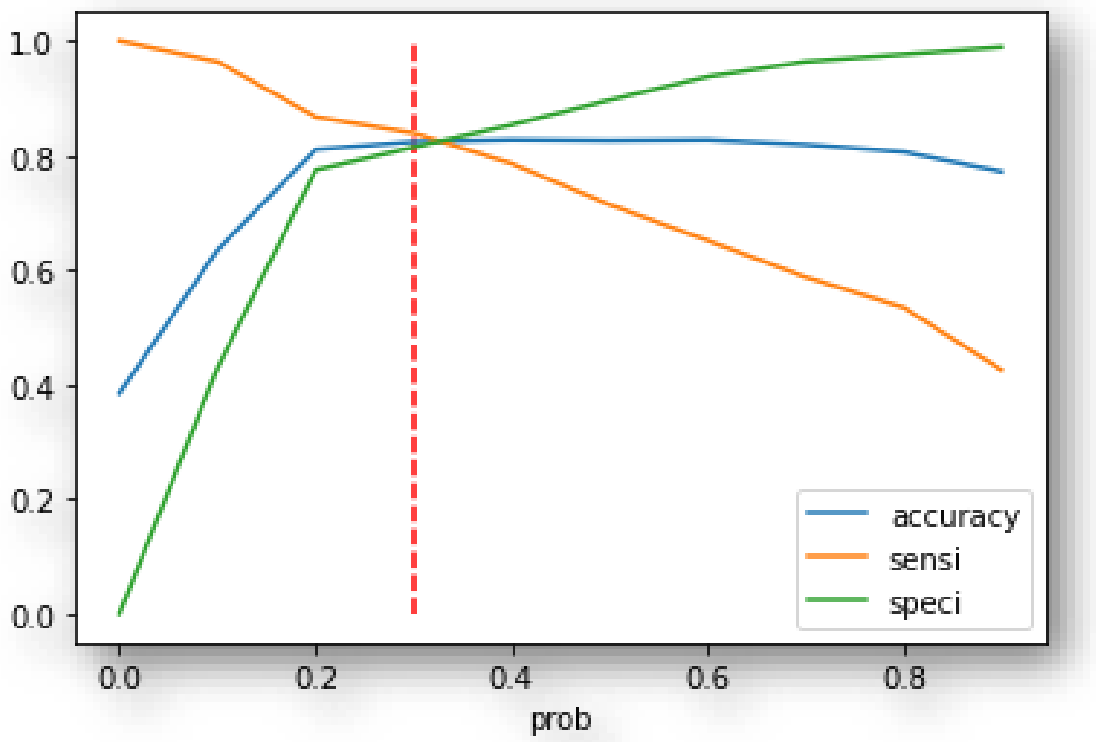
- As we can see ROC curve is quite good, it is covering around 90% area.



Now let's calculate accuracy sensitivity and specificity for various probability cutoffs.

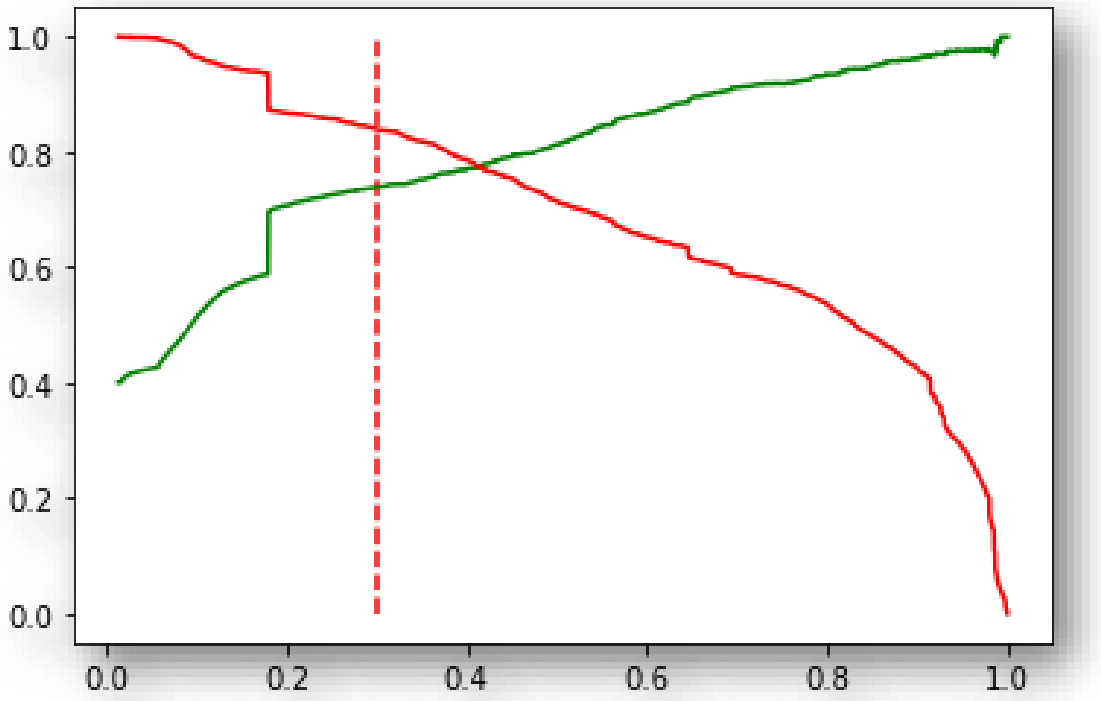
	prob	accuracy	sensi	speci
0.0	0.0	0.385136	1.000000	0.000000
0.1	0.1	0.635490	0.964023	0.429706
0.2	0.2	0.810581	0.867539	0.774904
0.3	0.3	0.824122	0.839330	0.814597
0.4	0.4	0.827744	0.786182	0.853777
0.5	0.5	0.826484	0.714227	0.896799
0.6	0.6	0.827586	0.651676	0.937772
0.7	0.7	0.819241	0.587899	0.964149
0.8	0.8	0.806487	0.534751	0.976697
0.9	0.9	0.772319	0.425593	0.989501

Plot accuracy sensitivity and specificity for various probabilities.



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

Precision Recall Trade off



- If we wanted a balanced model we might have gone for 0.42 or something but we need to get as much hot leads as possible.

Let's go with a higher sensitivity, we can trade a little bit of specificity for this case as business does not requires.

We are proceeding with cut off as 0.25

11. Predictions on Test Set

	Prospect ID	Converted	Converted_prob	final_predicted	Lead Score
0	3271	0	0.107247	0	10.72
1	1490	1	0.988831	1	98.88
2	7936	0	0.092021	0	9.20
3	4216	1	0.930832	1	93.08
4	3830	0	0.078609	0	7.86

- This model is trained with cutoff as 0.3 .

12. Model Evaluation on Test Set

Metrics	value
Accuracy	82.59%
Specificity	81.60%
Sensitivity	84.37%
Precision	72.33%
Recall	84.37%

- We are able to get about 82% of correct positive values.

13. Generating Hot Leads

	Prospect ID	Converted	Converted_prob	final_predicted	Lead Score
1	1490	1	0.988831	1	98.88
3	4216	1	0.930832	1	93.08
5	1800	1	0.823395	1	82.34
8	4223	1	0.986690	1	98.67
14	2570	1	0.695810	1	69.58

Hot leads data frame, containing all leads which have more than 60 lead score

- We have around 912 leads out of 9240 leads. It will reduce sales team effort.
- Conversion rate for the hot lead data frame is more than 91%.

CONCLUSION

From our Analysis, we come to a conclusion that the following factors are very important for X-Education for Lead conversion:

- we have more than 84% conversion rate for the hot_leads model suggested. we can see that there is a huge difference in number of leads sales team have to work on. this way there productivity will increase as well.
- Lead Source_Welingak Website, Lead Origin_Lead Add Form, Occupation_Working Professional and Total Time Spent on Website and Lead Source_Olark Chat are those variables which affect the model positively, and company should focus to improve these features in future.
- Rest features are negatively affecting the model. Company should try to focus on minimizing effects of these variables as they are negatively affecting the lead score.

	score (coeff_)
Lead Source_Welingak Website	3.661972
Lead Origin_Lead Add Form	2.702024
const	2.586070
Occupation_Working Professional	1.842879
Total Time Spent on Website	1.117845
Lead Source_Olark Chat	0.929907
Lead Origin_Landing Page Submission	-0.356571
Lead Source_Referral Sites	-0.937645
Do Not Email	-1.294756
Lead Quality_Low in Relevance	-1.706268
Lead Quality_Might be	-1.926080
Lead Quality_Not Sure	-4.044587
Lead Quality_Worst	-6.593489

Thank You