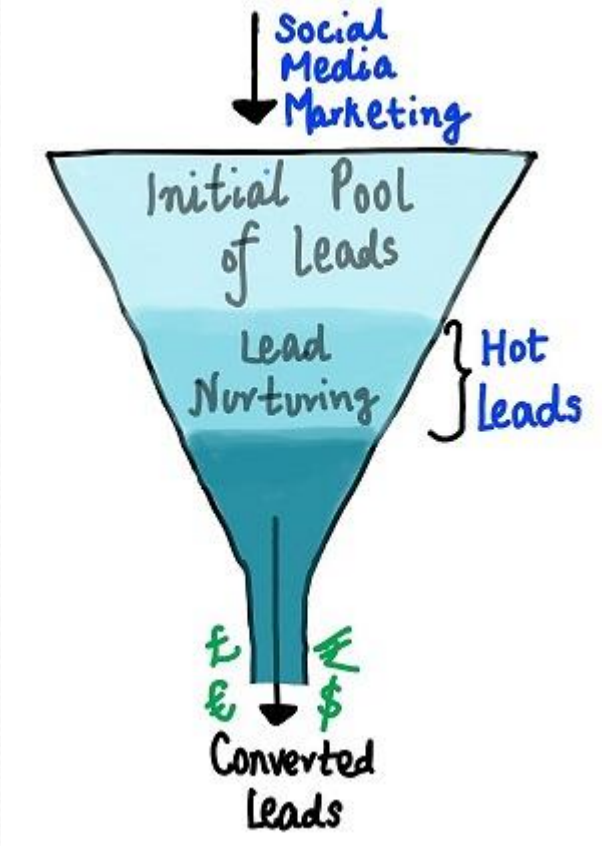
Lead Scoring

Case Study

Problem Statement

- X Education is online learning platform which sells courses through online. As most of professionals who are interested in the courses comes through online (generally it can be referred as leads), so its is always necessary for X Education to identify the right professionals who are most likely to buy their courses . They are calling these as hot leads.
- But Company is observing that, its lead conversion is rate is very poor and it reduced to 30%. That means out of 100 leads only 30 are purchasing the courses . Although company getting lots of leads , but they are not able to convert them into its customer .
- Thus, is is very much important to efficient this process and identify the hot leads , which they can easily convert. If company can identify the Hot leads, then rather calling to every leads , they can only call the Hot leads, which eventually save its time and increase the conversion ratio.
- The company requires to build a model wherein it needs 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 has given a ballpark of the target lead conversion rate to be around 80%.

As we can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, we need to nurture the potential leads well in order to get a higher lead conversion.



Lead Conversion Process - Demonstrated as a funnel

- **Finding the Solution:**

There are quite a few goals for this case study.
1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Approach and Steps taken for the Case Study:

From the above problem description, we concluded that this is a classification problem, hence we chose logistic Regression to calculate the Lead rate.

Below are the steps followed to solve this problem:

- **Data Reading and Understanding:**

  Here we tried to get the look and feel of the data, we observed following things:

  - Number of rows and columns
  - Data types of each columns
  - First few rows and how the data looks
  - Statistical aspects of the numerical features given in the data.
  - Understood the features given in the dataset using data dictionary excel file.

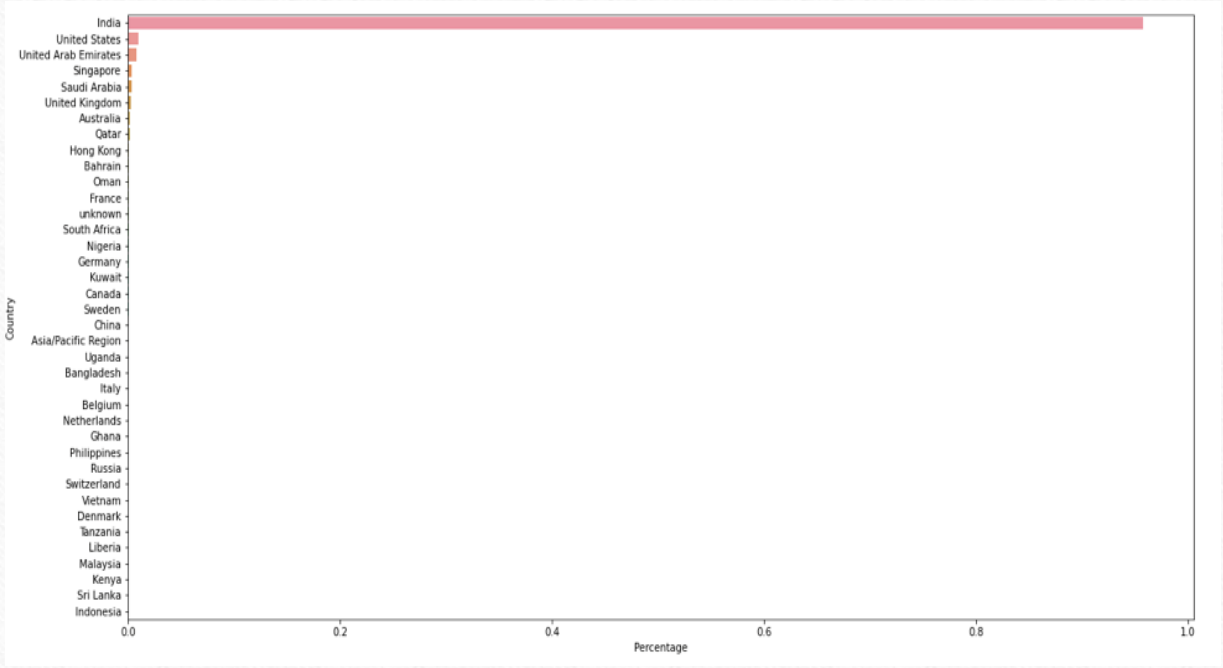
- **Data Cleaning and preparation:**
  - Here we checked for discrepancies in the dataset
    - Checking for any column names correction
    - Checking for null values and imputing them with appropriate methods
      - ✓ We used mode imputation for categorical columns.
      - ✓ We used mean imputation for numerical columns, if there is no skewness in data.
  - We used median imputation for numerical columns, if there is skewness in the data.
  - We have replaced some ”Select” values with NULL values as those are Nulls as per problem statement and business understanding .
  - We have converted some binary variables (Yes/No) to 0/1.
  - For categorical variables with multiple levels, created dummy features (one-hot encoded).

Missing Data Info:

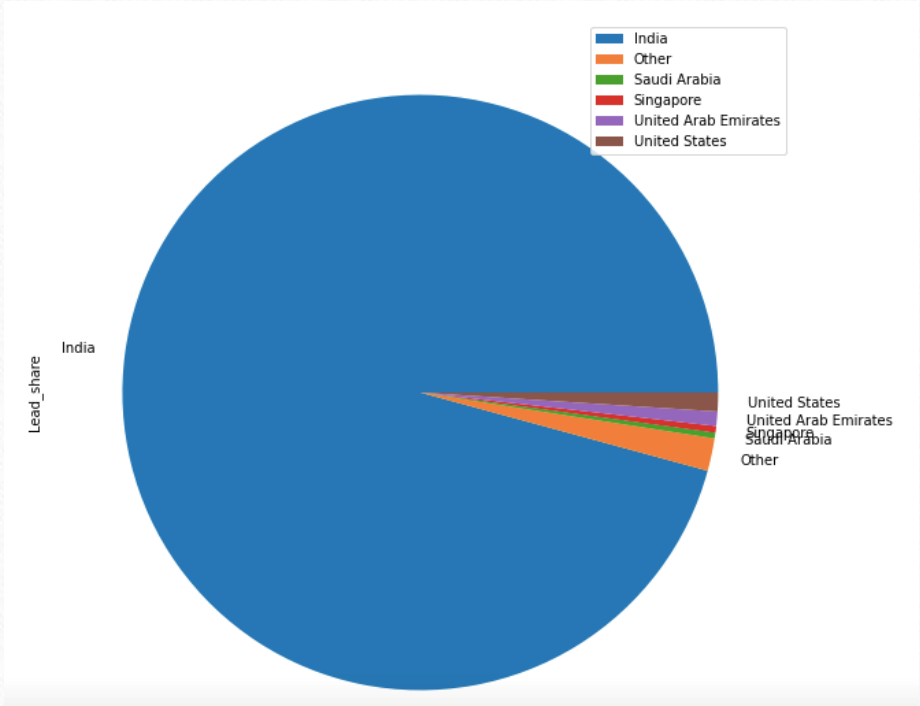
	Column_name	percentage_missing
0	Lead Number	0.00
1	Lead Origin	0.00
2	Lead Source	0.39
3	Do Not Email	0.00
4	Do Not Call	0.00
5	Converted	0.00
6	TotalVisits	1.48
7	Total Time Spent on Website	0.00
8	Page Views Per Visit	1.48
9	Last Activity	1.11
10	Country	26.63
11	Specialization	36.58
12	How did you hear about X Education	78.46
13	What is your current occupation	29.11
14	What matters most to you in choosing a course	29.32
15	Search	0.00
16	Newspaper Article	0.00
17	X Education Forums	0.00
18	Newspaper	0.00
19	Digital Advertisement	0.00
20	Through Recommendations	0.00
21	Tags	36.29
22	Lead Quality	51.59
23	Lead Profile	74.19
24	City	39.71
25	Asymmetrique Activity Index	45.65
26	Asymmetrique Profile Index	45.65
27	Asymmetrique Activity Score	45.65
28	Asymmetrique Profile Score	45.65
29	A free copy of Mastering The Interview	0.00
30	Last Notable Activity	0.00

Considering 60% threshold criteria there are two features which can be dropped that is Lead Profile and How did you hear about X Education. But both seems like a significant feature so keeping that for now.

- **Data Cleaning and preparation (Country Column) :**
  - During analysis we observed that majority of the countries have very less share, thus tagging everything below share 0.002 into others and below is the pie plot can be seen from the data.

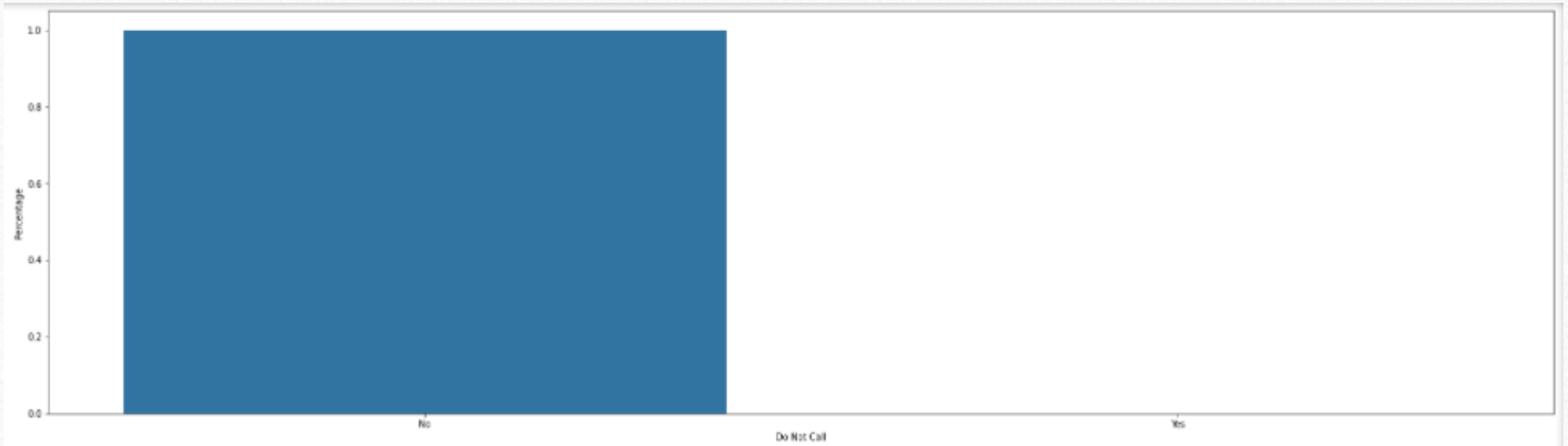


Before Cleaning



After Cleaning

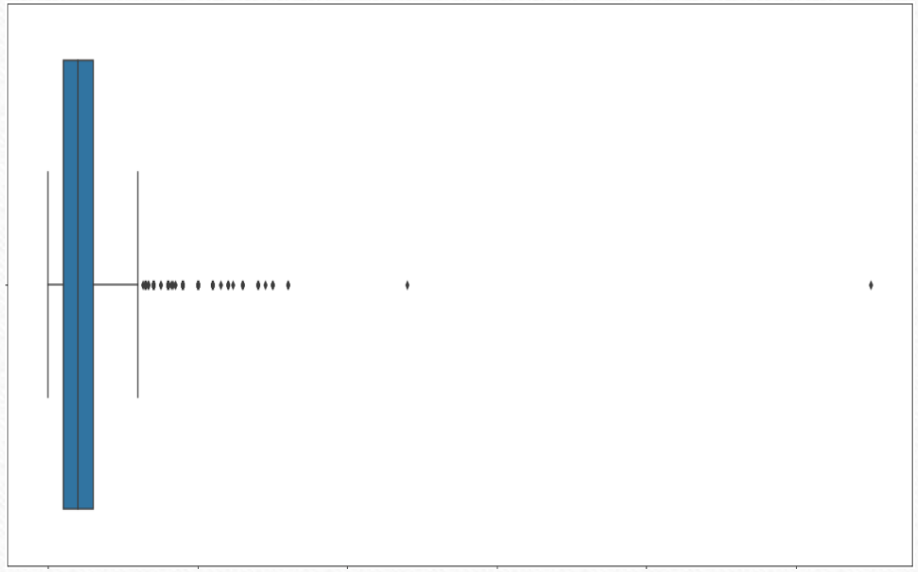
percentage share of each value for every categorical feature:



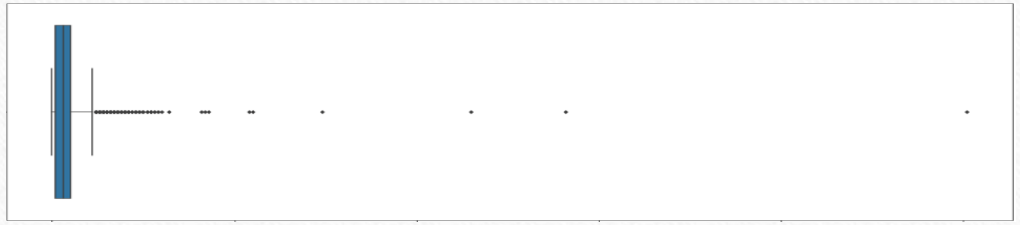
value share for Do Not Call

value share for What matters most to you in choosing a course

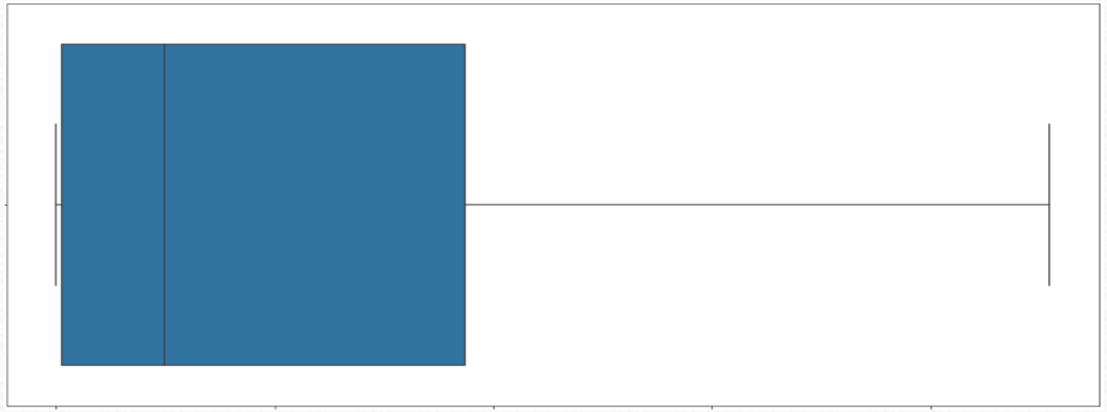
- **Checking Outliers:**
  - Outliers can be seen in features like 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'.
  - But as this is coming from actual data therefore dropping, normalizing or replacing these values won't be wise. these might result in actual impact towards Conversion



Total Visit



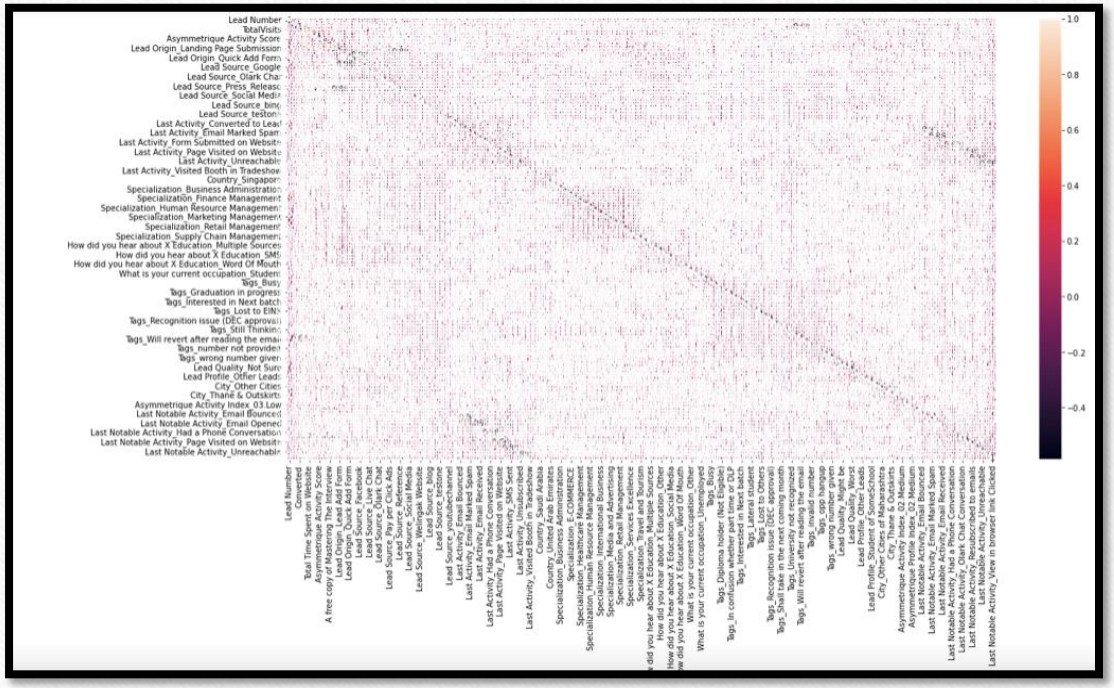
Page View per visit

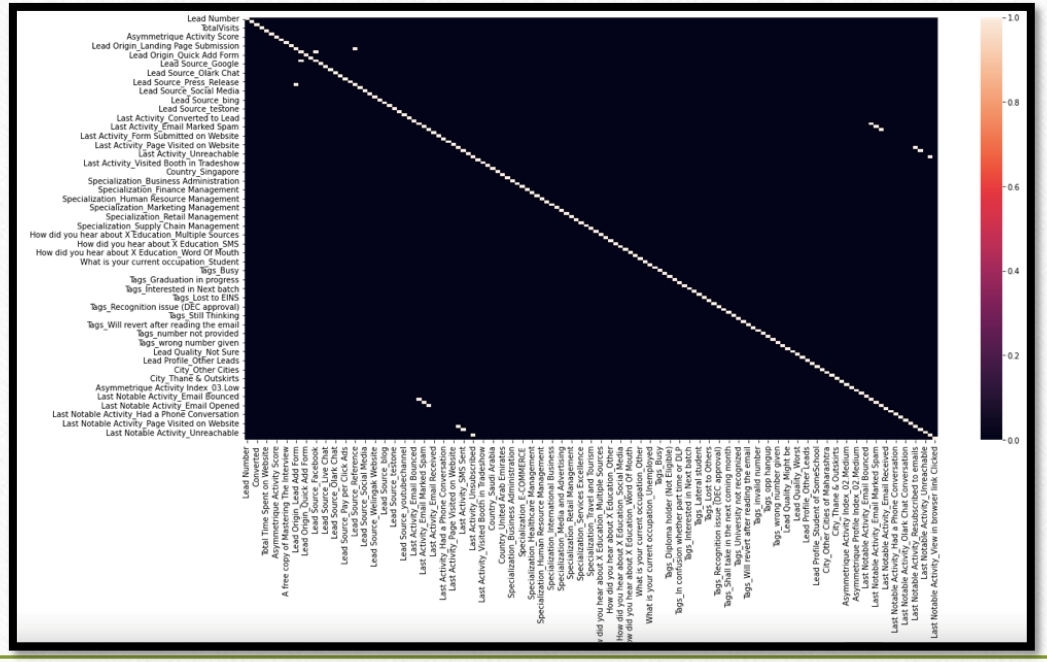


Total Time Spent on Website

- **Finding Correlations:**

Correlation to see if some features have very high correlation amongst themselves which might lead into overfitting, through we will use VIF while model fitting, but this is just to have an idea.





We can see there are high correlative relationships between features , but we can leave them for now and handle them with VIF in our further process.

- **Model Building :**
  - First, We split the data into training and testing sets. And first basic step for logistic regression is performing a train-test split . And we have chosen a ratio 70:30
  - We have used RFE for Feature Selection. And we ran **RFE with 20 variables** as output.
  - For all features VIF seems in check considering <5 criteria, therefore using p-value obtained in model statistics implies insignificant features, therefore we will try to delete those and see the impact.
  - After fourth iteration training accuracy appears to be too good [0.922] , so we should look into test accuracy and might have to adjust the cut-off as business problem. Another reason might be that initial conversion ratio itself was 38%. Let's continue with model validation.

Feature Selection Using RFE:

```
 ('Lead Source_Press_Release', False, 84),
 ('Lead Source_Reference', False, 27),
 ('Lead Source_Referral Sites', False, 59),
 ('Lead Source_Social Media', False, 97),
 ('Lead Source_WeLearn', False, 111),
 ('Lead Source_Welingak Website', True, 1),
 ('Lead Source_bing', False, 69),
 ('Lead Source_blog', False, 92),
 ('Lead Source_google', False, 64),
 ('Lead Source_testone', False, 119),
 ('Lead Source_welearnblog_Home', False, 96),
 ('Lead Source_youtubechannel', False, 100),
 ('Last Activity_Converted to Lead', False, 19),
 ('Last Activity_Email Bounced', False, 21),
 ('Last Activity_Email Link Clicked', False, 48),
 ('Last Activity_Email Marked Spam', False, 114),
 ('Last Activity_Email Opened', False, 98),
 ('Last Activity_Email Received', False, 102),
 ('Last Activity_Form Submitted on Website', False, 49),
 ('Last Activity_Had a Phone Conversation', False, 73),
```

```
[('Do Not Email', False, 9),
 ('TotalVisits', False, 81),
 ('Total Time Spent on Website', False, 7),
 ('Page Views Per Visit', False, 105),
 ('Asymmetrique Activity Score', False, 26),
 ('Asymmetrique Profile Score', False, 56),
 ('A free copy of Mastering The Interview', False, 99),
 ('Lead Origin_Landing Page Submission', False, 6),
 ('Lead Origin_Lead Add Form', False, 10),
 ('Lead Origin_Lead Import', False, 62),
 ('Lead Origin_Quick Add Form', False, 82),
 ('Lead Source_Direct Traffic', False, 67),
 ('Lead Source_Facebook', False, 43),
 ('Lead Source_Google', False, 70),
 ('Lead Source_Live Chat', False, 118),
 ('Lead Source_NC_EDM', False, 11),
 ('Lead Source_Olark Chat', False, 41),
 ('Lead Source_Organic Search', False, 68),
 ('Lead Source_Pay per Click Ads', False, 113),
```

Model building and backward elimination with statsmodel

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6
Model:	GLM	**Df Residuals:**	6
Model Family:	Binomial	**Df Model:**	
Link Function:	Logit	**Scale:**	1.0
Method:	IRLS	**Log-Likelihood:**	-133
Date:	Sat, 12 Nov 2022	**Deviance:**	266
Time:	19:03:20	**Pearson chi2:**	1.80e-
No. Iterations:	23	**Pseudo R-squ. (CS):**	0.6
Covariance Type:	nonrobust		

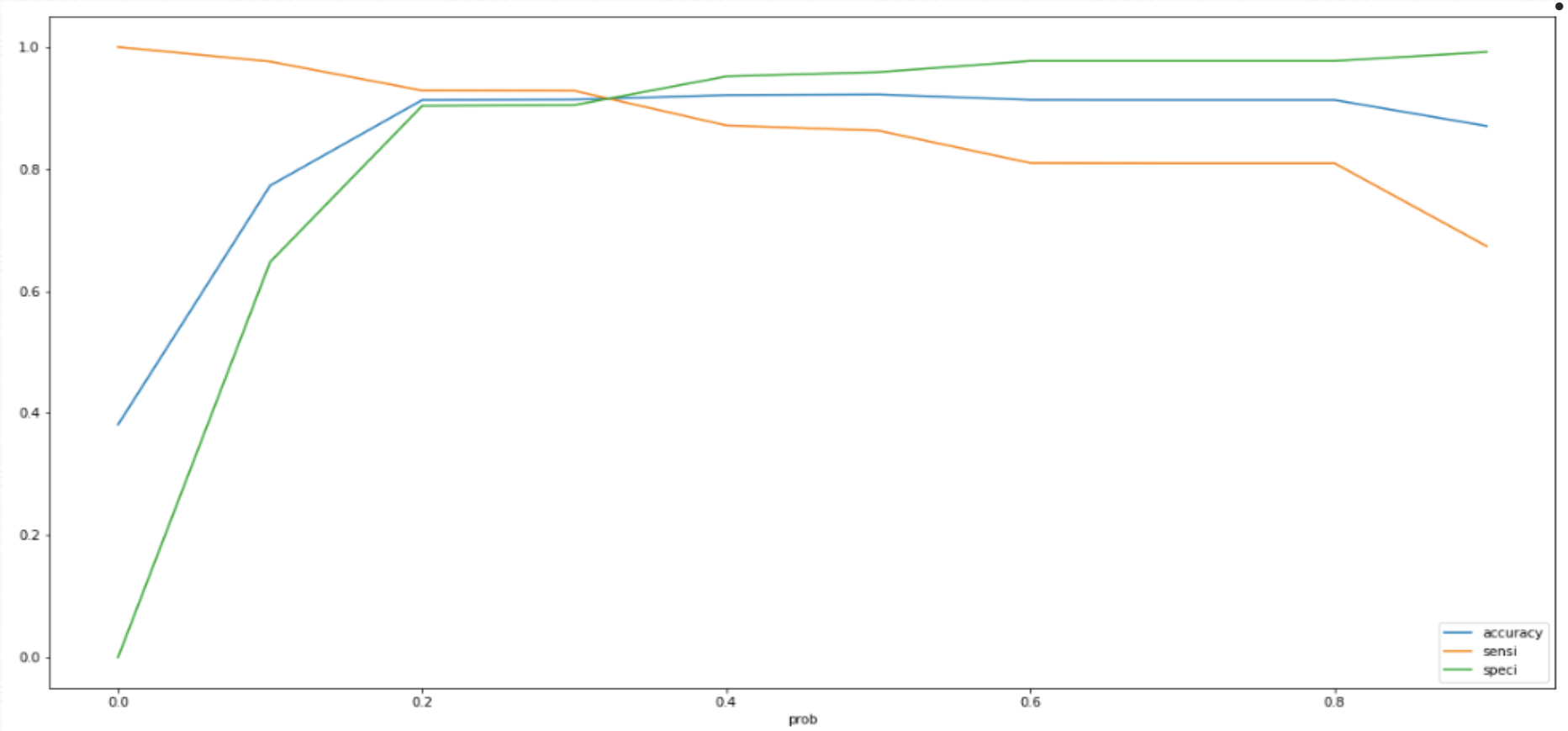
Lead Source_Welingak Website	3.6895	0.751	4.912	0.000	2.217	5.162
Last Activity_SMS Sent	2.1721	0.114	19.024	0.000	1.948	2.396
What is your current occupation_Unemployed	1.6906	0.114	14.813	0.000	1.467	1.914
What is your current occupation_Working Professional	1.9489	0.331	5.880	0.000	1.299	2.598
Tags_Closed by Horizzon	7.0996	1.012	7.015	0.000	5.116	9.083
Tags_Diploma holder (Not Eligible)	-23.1531	1.73e+04	-0.001	0.999	-3.4e+04	3.4e+04
Tags_Interested in full time MBA	-2.3706	0.736	-3.221	0.001	-3.813	-0.928
Tags_Interested in other courses	-2.3569	0.336	-7.013	0.000	-3.016	-1.698
Tags_Lost to EINS	7.0025	0.818	8.557	0.000	5.399	8.606
Tags_Not doing further education	-3.3805	1.044	-3.237	0.001	-5.427	-1.334
Tags_Ringing	-4.3673	0.235	-18.616	0.000	-4.827	-3.907
Tags_Will revert after reading the email	3.8713	0.189	20.519	0.000	3.502	4.241
Tags_invalid number	-4.6199	1.031	-4.479	0.000	-6.642	-2.598
Tags_number not provided	-24.8545	2.48e+04	-0.001	0.999	-4.86e+04	4.86e+04
Tags_opp hangup	-2.8657	0.806	-3.555	0.000	-4.445	-1.286
Tags_switched off	-4.8085	0.524	-9.174	0.000	-5.836	-3.781
Tags_wrong number given	-25.0255	2.08e+04	-0.001	0.999	-4.09e+04	4.08e+04
Lead Quality_Worst	-3.3311	0.547	-6.091	0.000	-4.403	-2.259

- **Model Validation :**
  - We could find out the sensitivity of our logistic regression model as,
    - Sensitivity: 0.8629359286293593 ,
    - Specificity: 0.9587706146926537
    - TPR: 0.9280418665503707
    - TNR: 0.9190419161676646
  - As our target is to identify Hot Leads which means we should not be losing out on customers who had high chances of converting but because we were not able to tag them as hot leads will leads to actual loss. Then means Recall = TP/TP+FN should be increased as much as possible.

    TP+FN is total possible hot leads; TP is hot leads tagged correctly.

    As recall = Sensitivity and that is close to 86% as per 0.5 cut-off which means we are able to increase conversion rate to 86% by tagging 86% of all possible hot leads correctly.

- **Create plot of accuracy , sensitivity and specificity for various probabilities :**

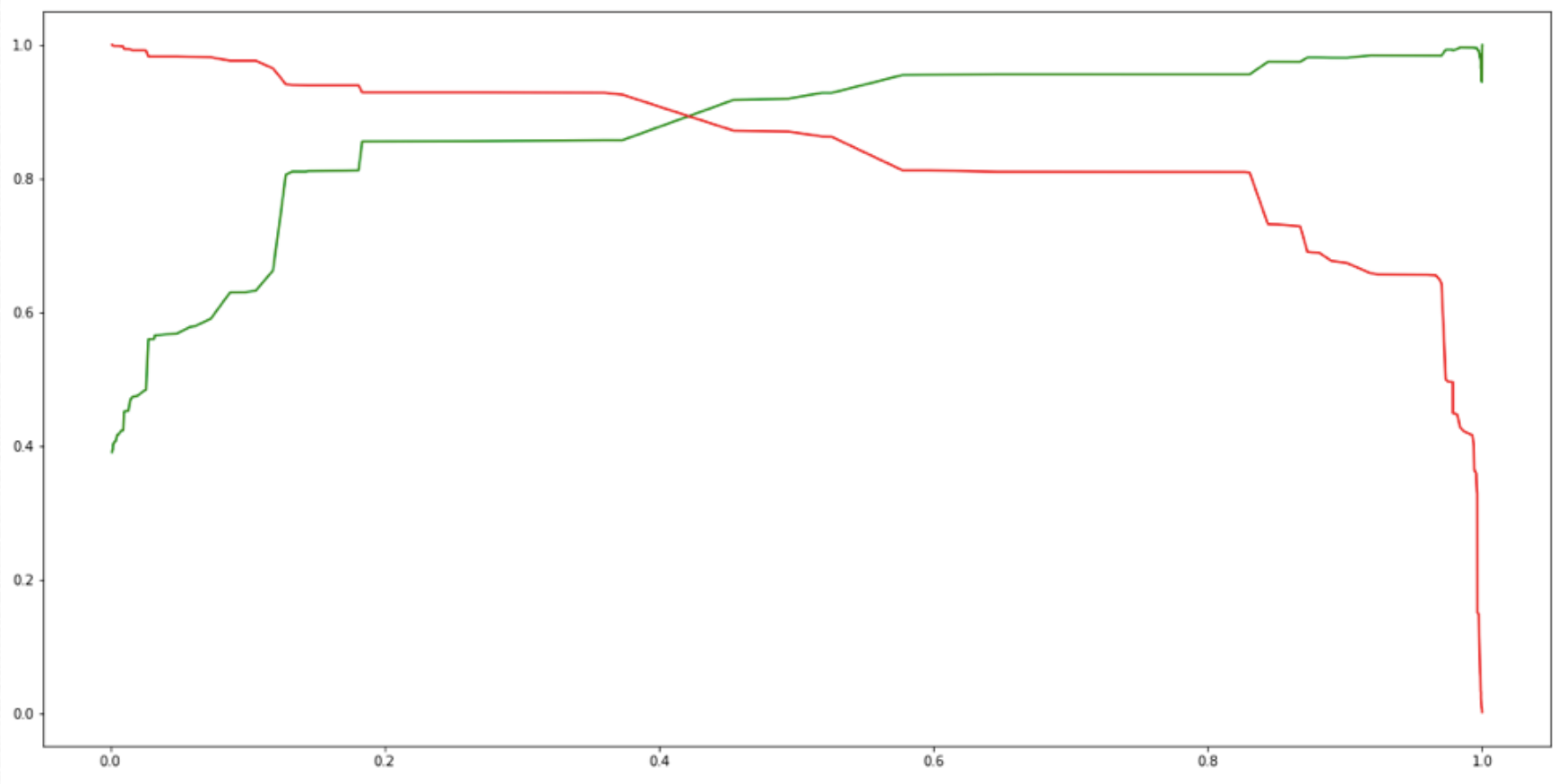


- X axis = Probabilities, Y axis = accuracy , sensitivity and specificity

- **From the above curve above, 0.35 is the optimum point to take it as a cutoff probability. Contd..**

- **After getting optimal cut-off our conversion rate or correct detection of hot leads has increased to ~93%**

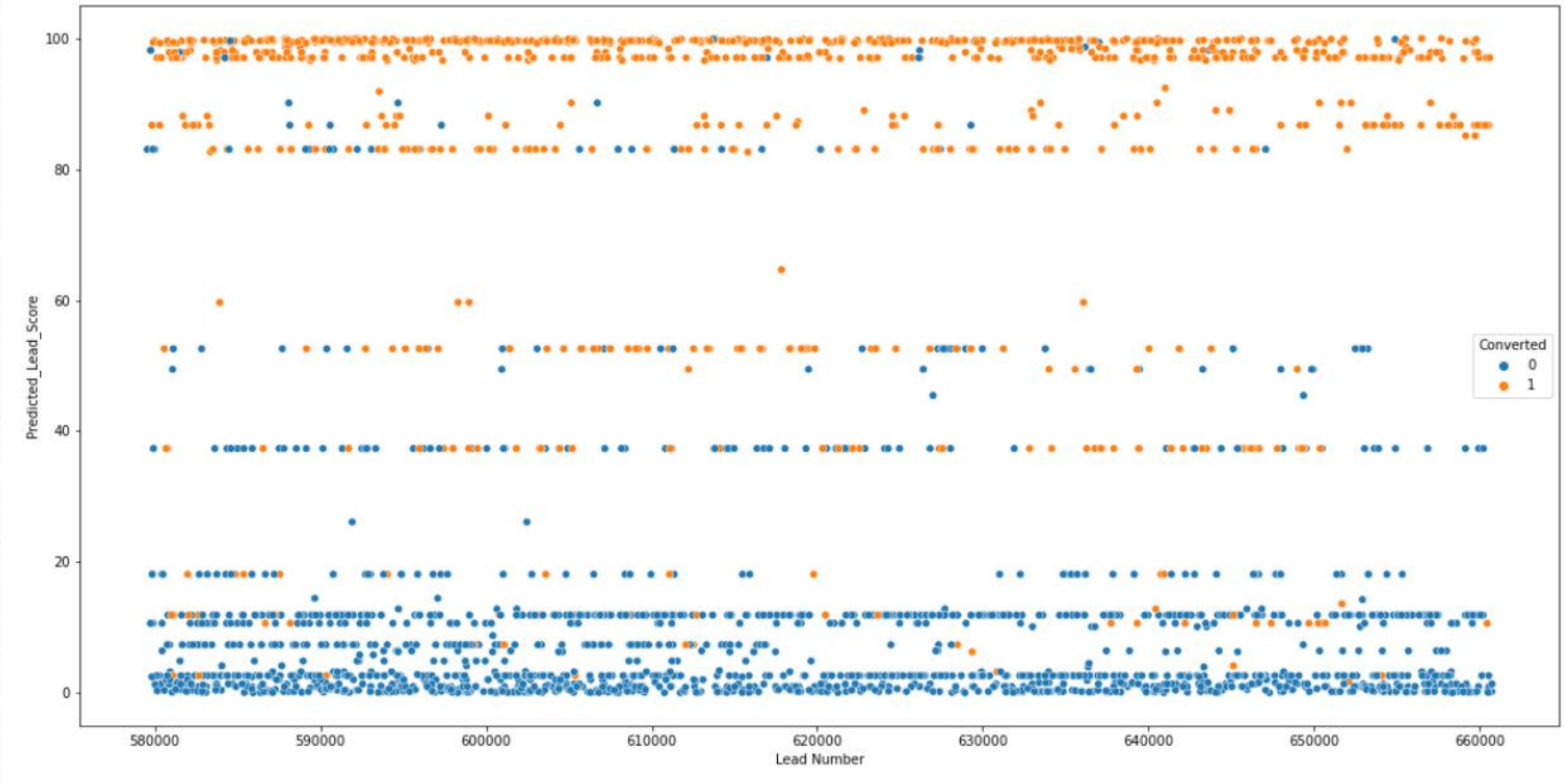
Precision-Recall Curve



- **Testing Results on Test Data and Comes to conclusions:**
  - **This seems like a great model to accurately predict as many hot leads as possible correctly. As we can see in train data, we are able to correctly identify the hot-leads by 93% and for test-data by 94%.**
  - **Also, as per model statistics, most significant 5 features impacting conversion are : Tags_Closed by Horizzon, Tags_Lost to EINS, Tags_switched off, Tags_invalid number and Tags_Ringing**
  - **Most significant 3 features with positive impact :**
    - **Tags_Closed by Horizzon,**
    - **Tags_Lost to EINS,**
    - **Tags_Will revert after reading the email**
  - **Most significant 3 features with negative impact :**
    - Tags_switched off,
    - Tags_invalid number,
    - Tags_Ringing
  - **That means of the prospects current status is Switched off, Invalid or Ringing that means there are high chances of non conversion,**
    However, if the status is Horizzon, EINS or will revert after reading the email that means there are high chances of conversion

- **At last, we use the model to predict on test data again to conduct a last business step showing most of the converted prospects should have high prediction lead score (0-100).**

Prediction Lead Score Tagged with actual Conversion tag



THANK YOU