LEADSSCORE CASE STUDY

- Nisar Killedar
- Nupur Dey
- Nithish

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%.
- 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. 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.

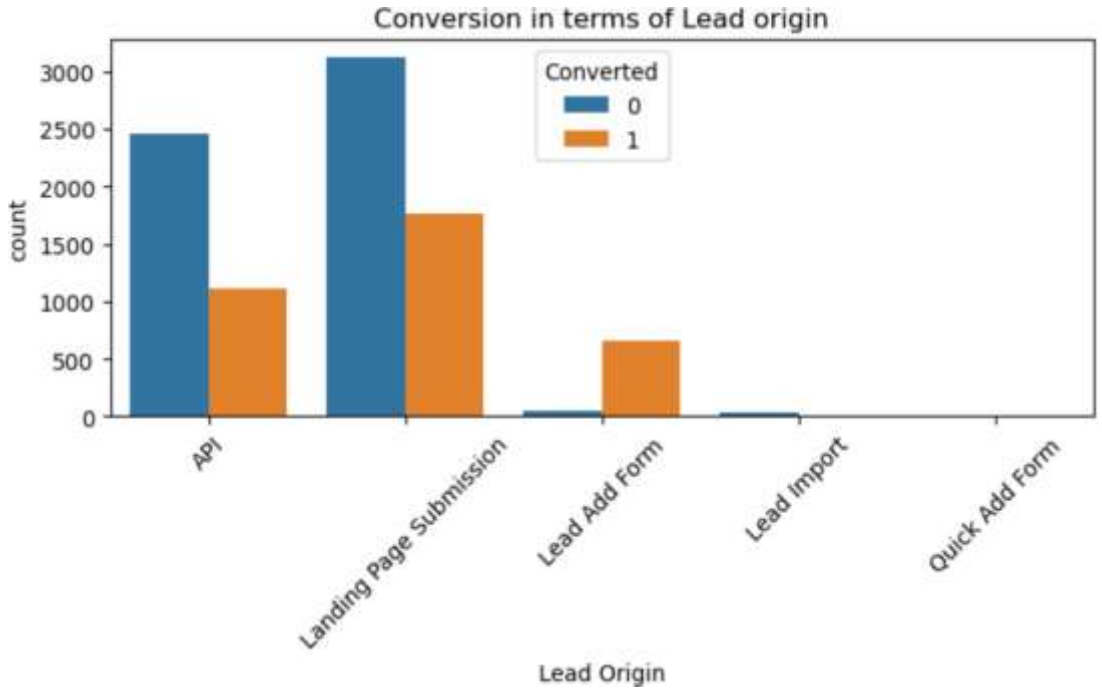
Goal of the Case Study:

- 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.
- 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.

SOLUTION METHODOLOGY

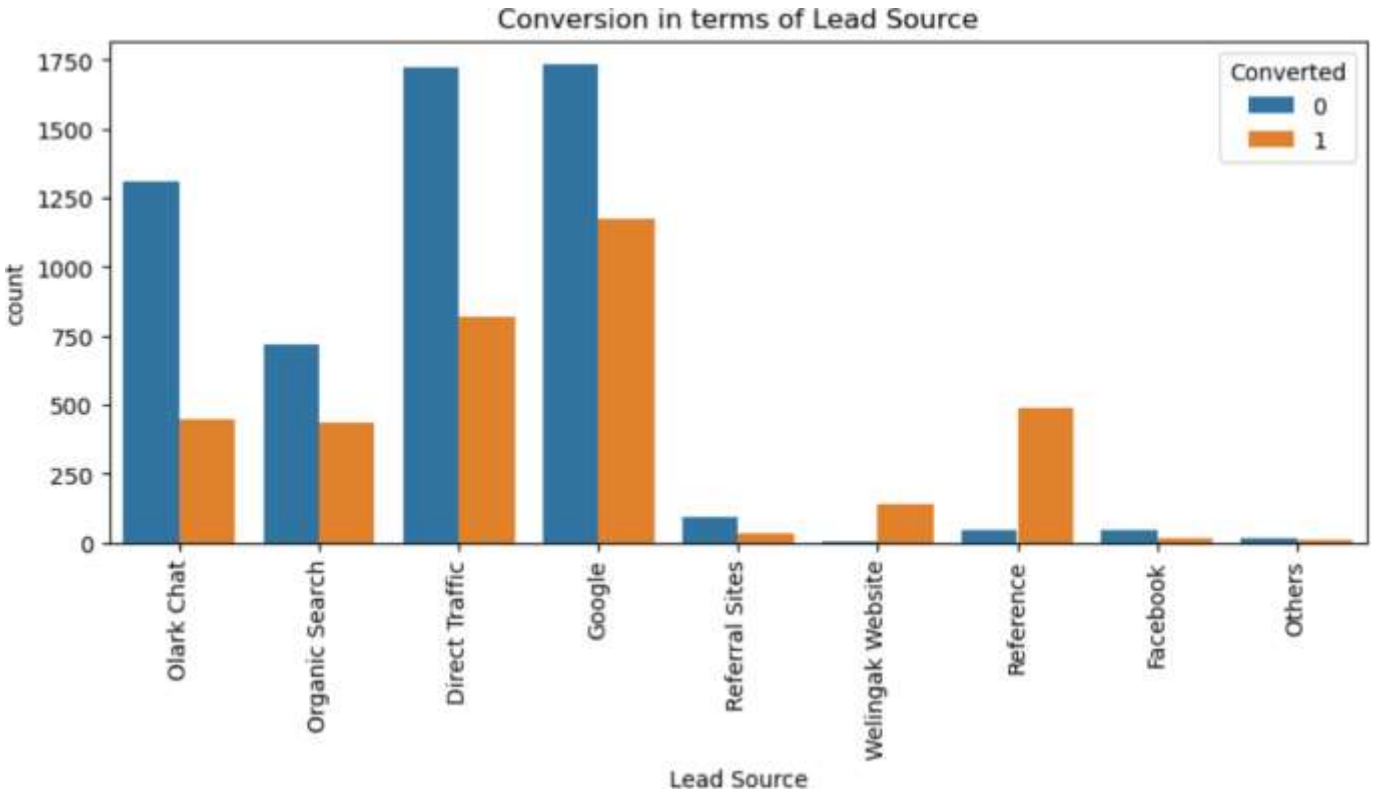
- Reading and understanding the data
- Data Cleaning
  - Delete high null values
  - Handle categorical null values
  - Handle numerical null values
- Exploratory Data Analysis
  - Data Imbalance check
  - Univariate analysis
- Data Preparation
  - Outlier treatment
  - Convert binary categories
  - Create dummy variables
  - Train Test Split
  - Feature Scaling
- Model Building
- Model Evaluation
- Conclusion

EDA Lead Origin



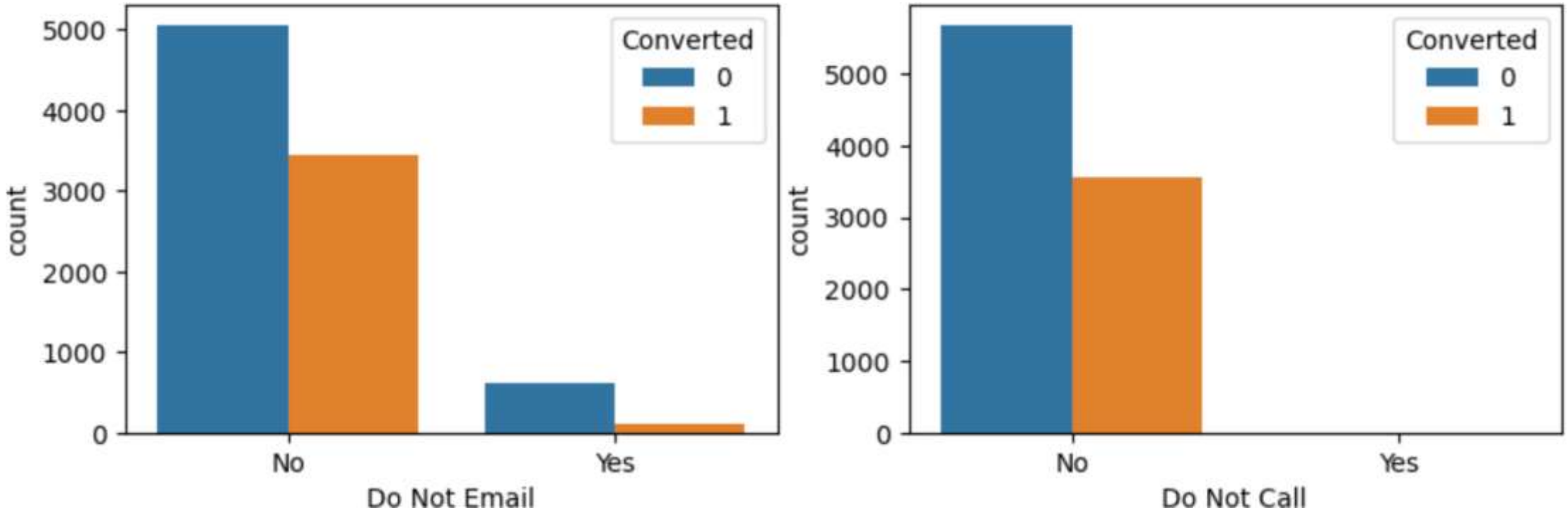
- Observation From the above plot and Lead origin conversion summary, we can infer that: Lead Add Form has the highest conversion rate at 92% API and Landing Page Submission have 31% and 36% conversion rate but they generate maximum leads counts. Lead Import has the least amount of conversions and leads count. To improve overall lead conversion rate, focus should be on improving lead conversion rate of API and Landing Page Submission. Also, generate more leads from Lead Add form since they have a very good conversion rate

EDA Lead Source



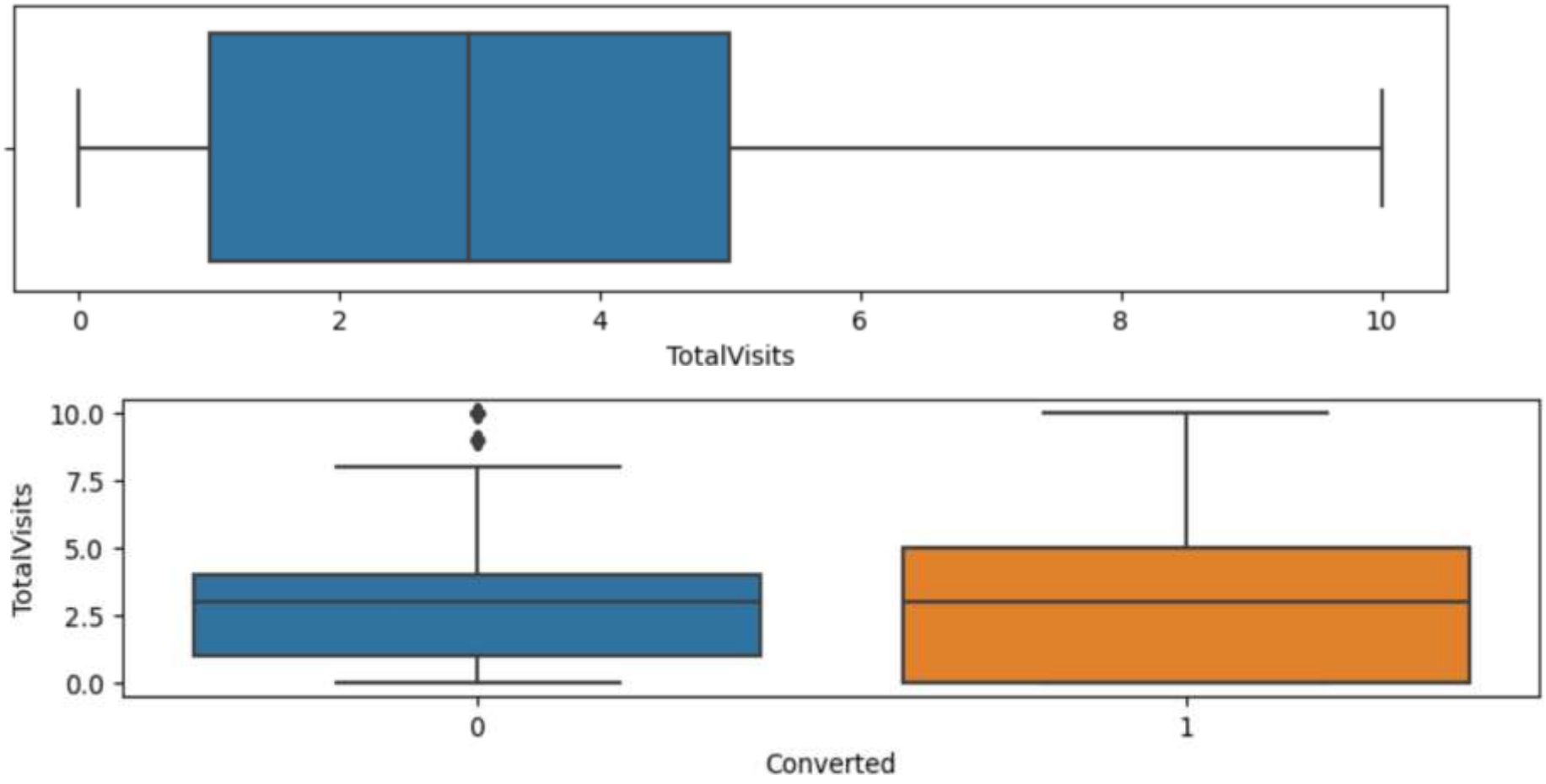
- Google and direct traffic generates maximum number of leads but has conversion rate of 40% and 32%.
- Welingak website and References has highest conversion ratesaround 98% and 93% but generates less number of leads.

Do Not Email & Do Not Call

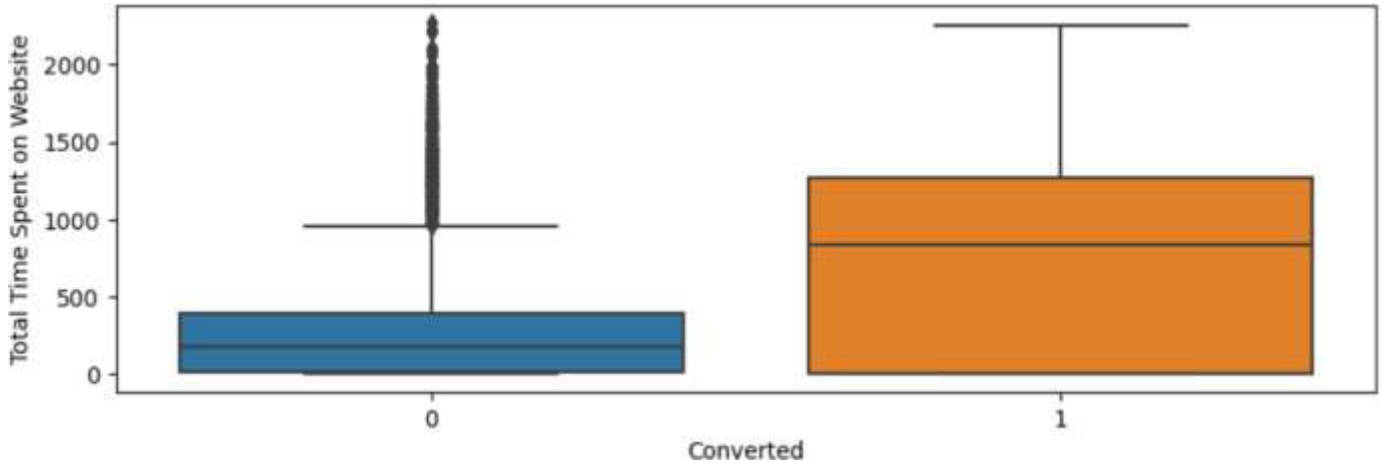


Most of the customers do not like to be called or receive emails about the course.

Total visits

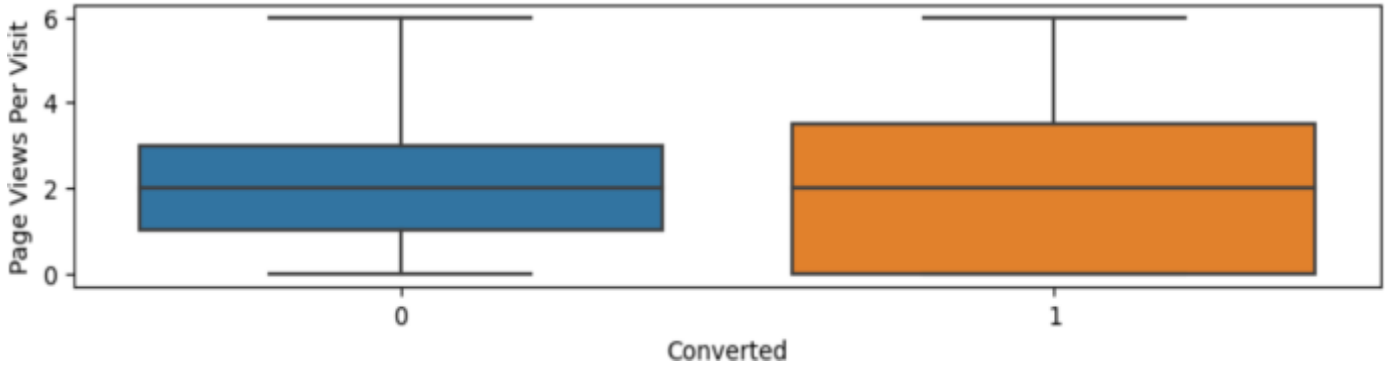


Total Time Spent on Website

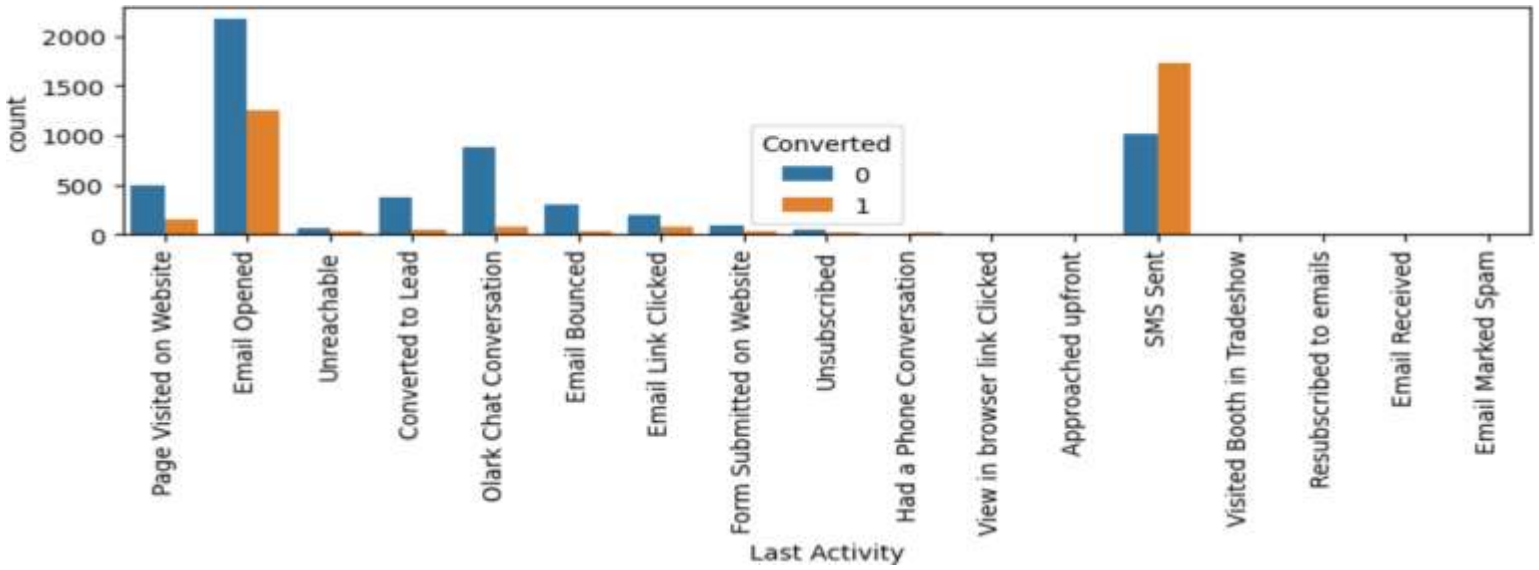


The total time spent on website is directly proportional to conversion rate

Page Views Per Visit



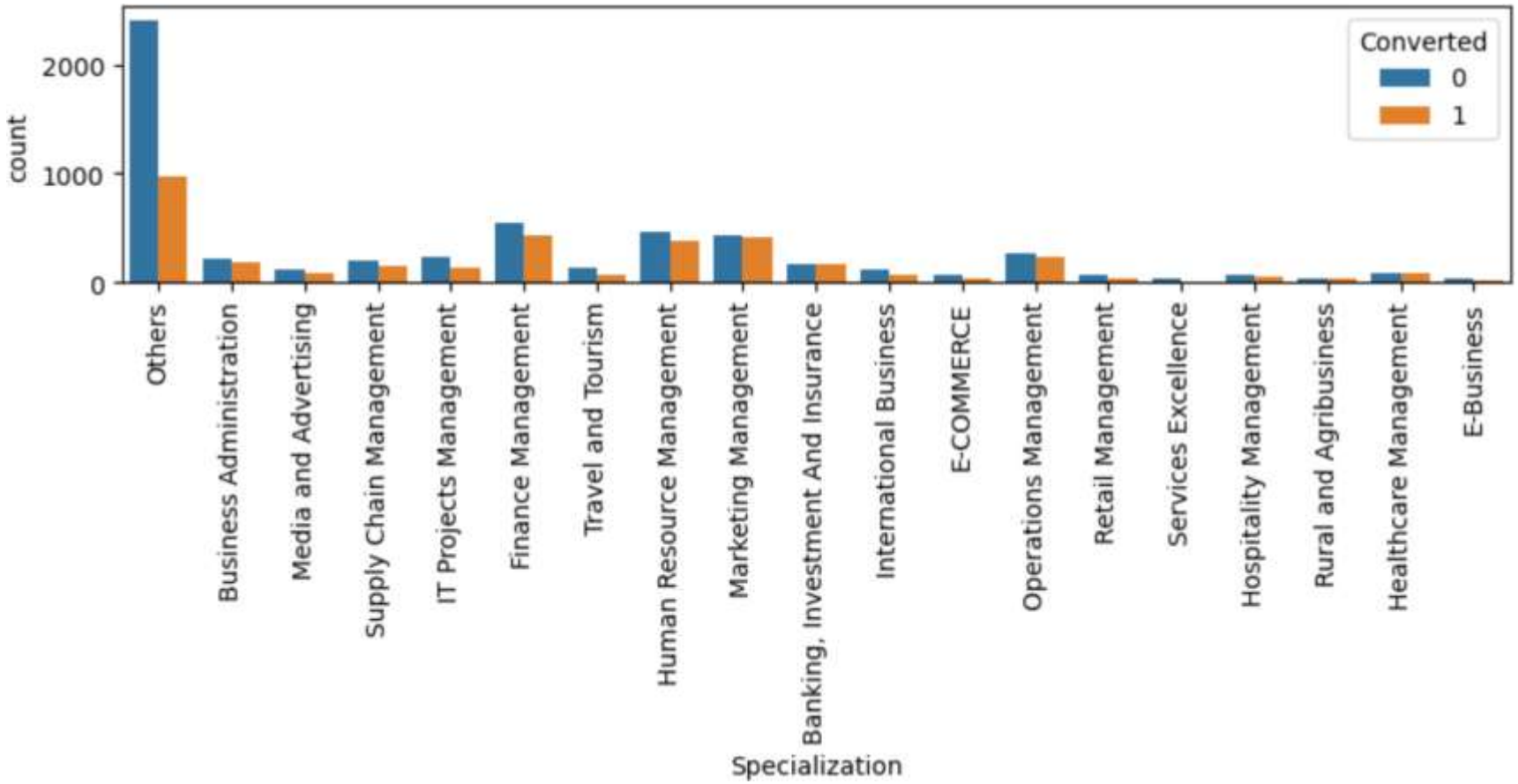
Last Activity



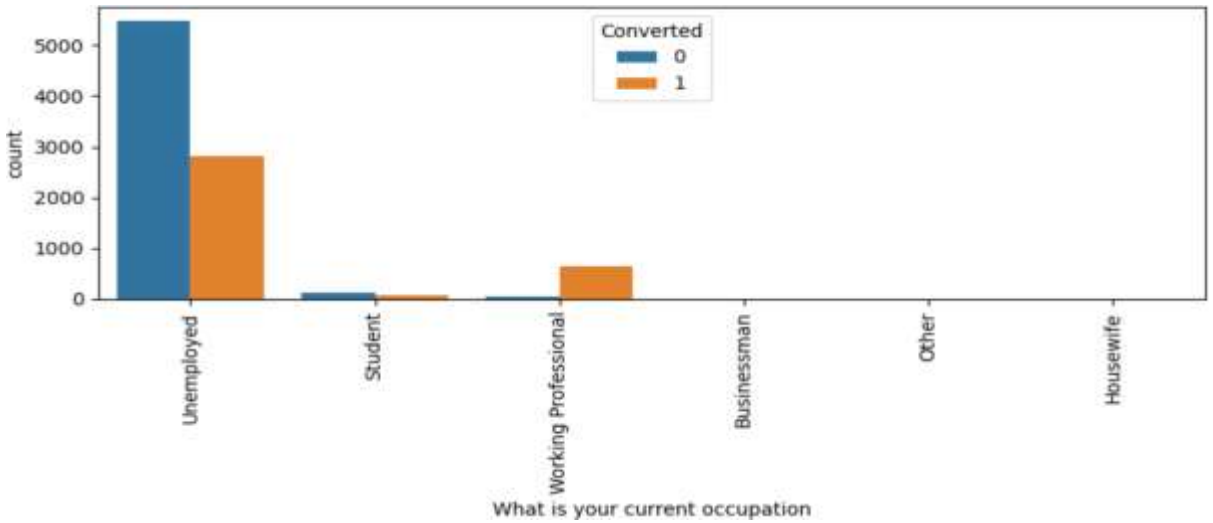
1. Maximum leads are generated from people with last activity - Email opened and SMS sent.
2. Conversion rate is around 63% and 36% .

3.To improve overall lead conversion rate, focus should be on improving lead conversion of people with last activity -olark chat conversation, SMS sent and Page Visited on Website .

Specialization

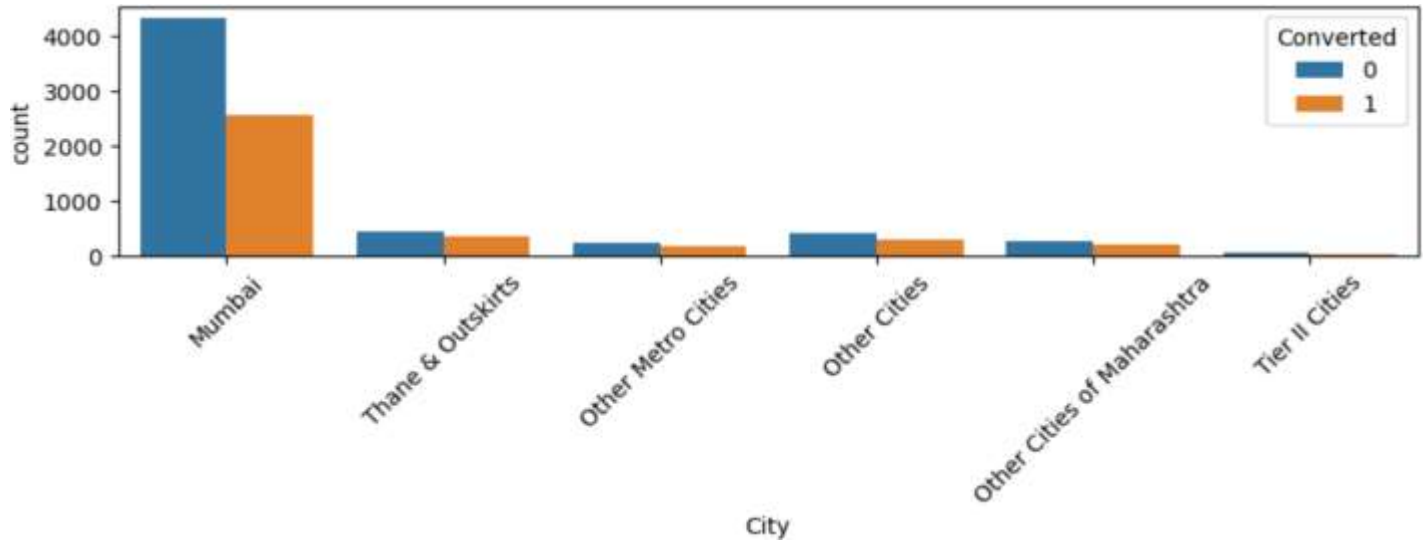


What is your current occupation ?



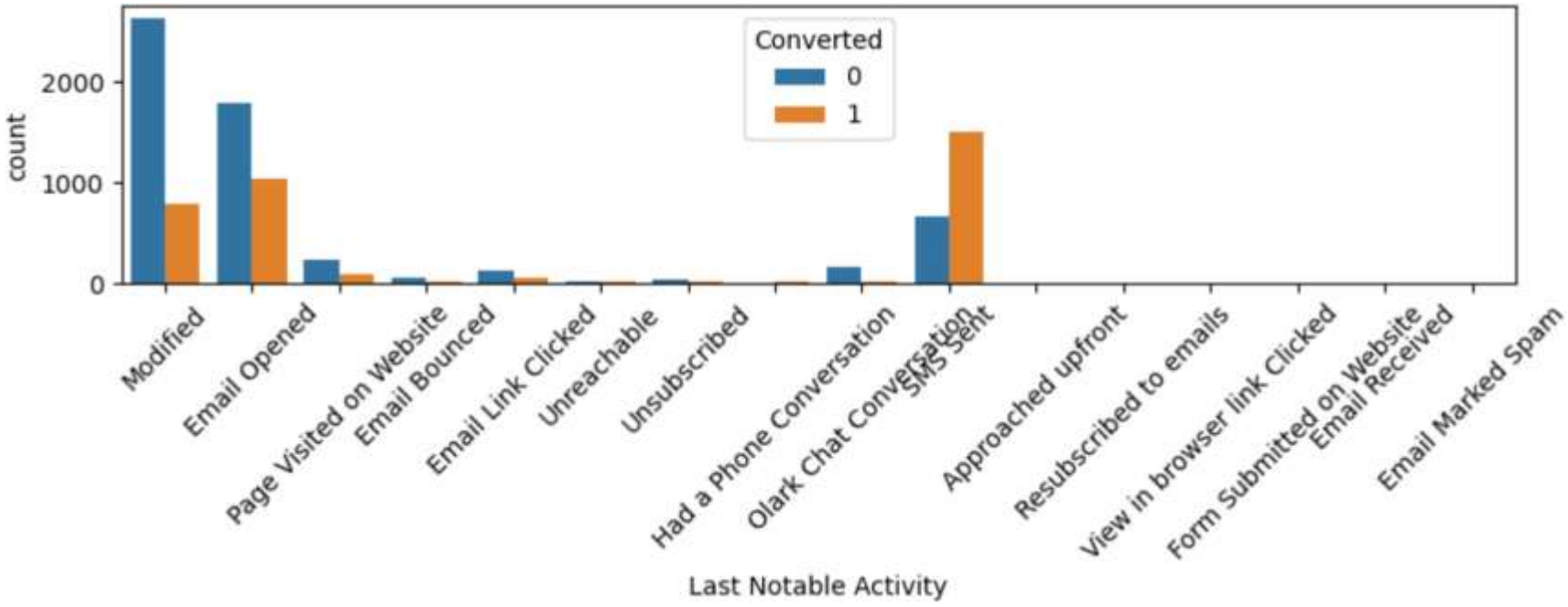
Working Professionals and Unemployed people generates maximum leads .

City



Maximum leads are generated from Mumbai city with conversion rate of around 36% .Hence focus should me more on increasing conversion rate of Mumbai city

Last Notable Activity



- Data Preparation
  - Create dummy variables for categorical variables
  - Drop original columns
- Train Test Split
  - Train size – 70% of the data
  - Test size – 30% of the data
- Feature Scaling
  - Using standard scalar we scale the below columns
  - Total Visits, Total Time Spent on Website, Page Views Per Visit
- Feature selection using RFE
  - Using RFE function we select 20 variables which would help the most for model building

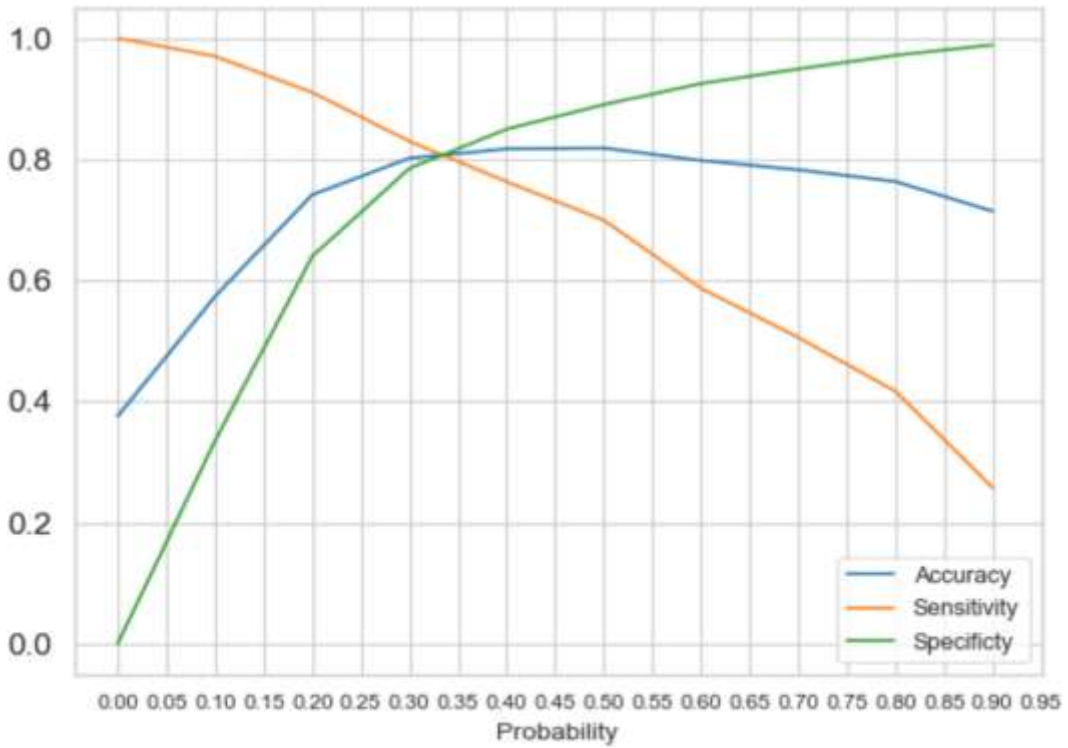
```
In [229]: col1 = X_train.columns[rfe.support_]
          col1

Out[229]: Index(['Do Not Email', 'Total Time Spent on Website',
                 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
                 'Lead Source_Welingak Website', 'Last Activity_Converted to Lead',
                 'Last Activity_Email Bounced', 'Last Activity_Had a Phone Conversation',
                 'Last Activity_Olark Chat Conversation',
                 'Last Activity_Page Visited on Website',
                 'What is your current occupation_Housewife',
                 'What is your current occupation_Student',
                 'What is your current occupation_Unemployed',
                 'What is your current occupation_Working Professional',
                 'Last Notable Activity_Email Link Clicked',
                 'Last Notable Activity_Email Opened',
                 'Last Notable Activity_Had a Phone Conversation',
                 'Last Notable Activity_Modified',
                 'Last Notable Activity_Olark Chat Conversation',
                 'Last Notable Activity_Unreachable'],
                dtype='object')
```

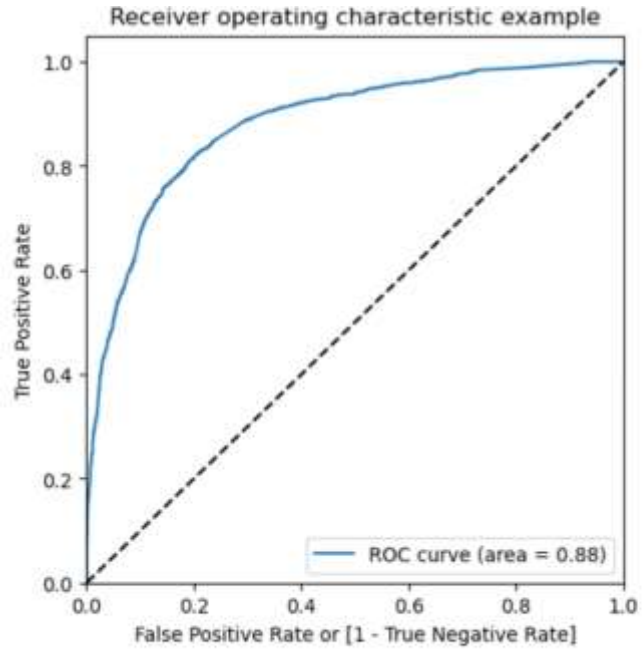
- Model Building
  - We use Generalized Linear Model to build the model
  - We use Variance Inflation Factor (VIF) and P value to drop the unwanted columns
  - After iterating few times we decide with set of columns to start predicting
- Once we are done with the column selection we can start predicting and evaluating the model
- We have the below results once we evaluate the model
  - Sensitivity : 70 %
  - Specificity : 89%

Out[254]:

	Features	VIF
7	Last Activity_Had a Phone Conversation	2.39
13	Last Notable Activity_Had a Phone Conversation	2.39
14	Last Notable Activity_Modified	2.07
8	Last Activity_Olark Chat Conversation	2.05
6	Last Activity_Email Bounced	1.83
0	Do Not Email	1.81
3	Lead Source_Olark Chat	1.67
2	Lead Origin_Lead Add Form	1.43
15	Last Notable Activity_Olark Chat Conversation	1.33
5	Last Activity_Converted to Lead	1.27
4	Lead Source_Welingak Website	1.25
1	Total Time Spent on Website	1.23
9	Last Activity_Page Visited on Website	1.13
10	What is your current occupation_Working Profes...	1.13
12	Last Notable Activity_Email Opened	1.11
11	Last Notable Activity_Email Link Clicked	1.02

- Plotting the ROC Curve
  - tells how much the model is capable of distinguishing between classes





The ROC Curve should be a value close to 1. We are getting a value of 0.88 indicating a good predictive model.

- Finding the optimal Cut off Point

- From the above curve we can see that the optimal cut-off is at 0.35. This is the point where all the parameters -Accuracy,Sensitivity,Specificity are equally balanced
- We will only consider data with probability > 0.35
- Once we have the data , we can start predicting and checking the accuracy
- Below are the results
    - Overall Accuracy : 81 %
    - Specificity : 81%
    - Sensitivity : 80%
- When we are selecting the optimal cut-off = 0.35, the various performance parameters Accuracy, Sensitivity & Specificity are all 80%

- Precision and Recall
    - precision and recall which tells us the score for result relevancy and how many truly relevant results are returned
    - Below are the results
        - Precision score: 72 %
        - Recall score: 80 %

- Making Predictions on test set
    - Once we test the model on test data below are the results:
        - Overall accuracy : 81%
        - Sensitivity : 80%
        - Specificity : 82%
        - Precision Score : 75%
        - Recall Score : 80%

Final Observation

- Lets compare the Model Performance parameters obtained for Train & Test data:
  - Train Data:
    - Accuracy : 81%
    - Sensitivity : 79%
    - Specificity : 81%
    - Precision score: 72%
    - Recall score : 80%
  - Test Data:
    - Accuracy : 81%
    - Sensitivity : 80%
    - Specificity : 82%
    - Precision Score: 75%
    - Recall Score : 80%
- We got around 1% difference on train and test data's performance metrics. This implies that our final model didn't overfit training data and is performing well.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.
- Depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.

Recommendations

1. The sales team of the X-Education should focus on the leads having lead origin - lead add form , occupation - Working Professional , Lead source - Welling Ak website.
2. Sales Team of the company should first focus on the 'Hot Leads'
3. The 'Cold Leads'(Customer having lead score <= 35) should be focused after the Sales Team is done with the 'Hot Leads'.
4. High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.
5. We have high recall score than precision score. Hence this model has an ability to adjust with the company's requirements in coming future.
6. Its better that we can ask students to fill many important fields like location details which can be used for futher analysis
7. We can ignore customers who do not want to be called about the course.
8. If the Last Notable Activity is Modified, he/she may not be the converted to lead