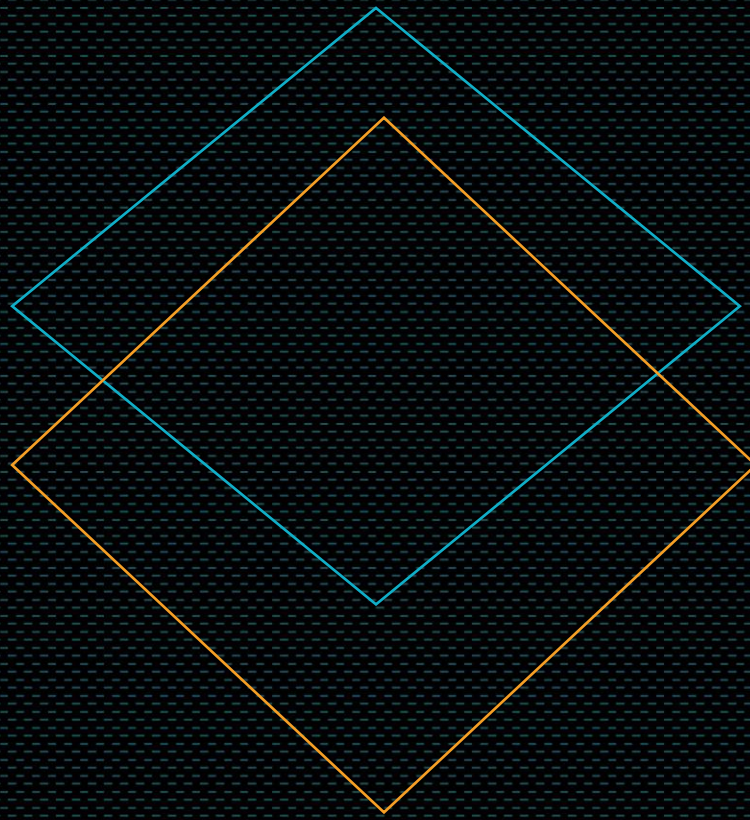


# Lead Scoring Case Study

# Introduction

## 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. 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.

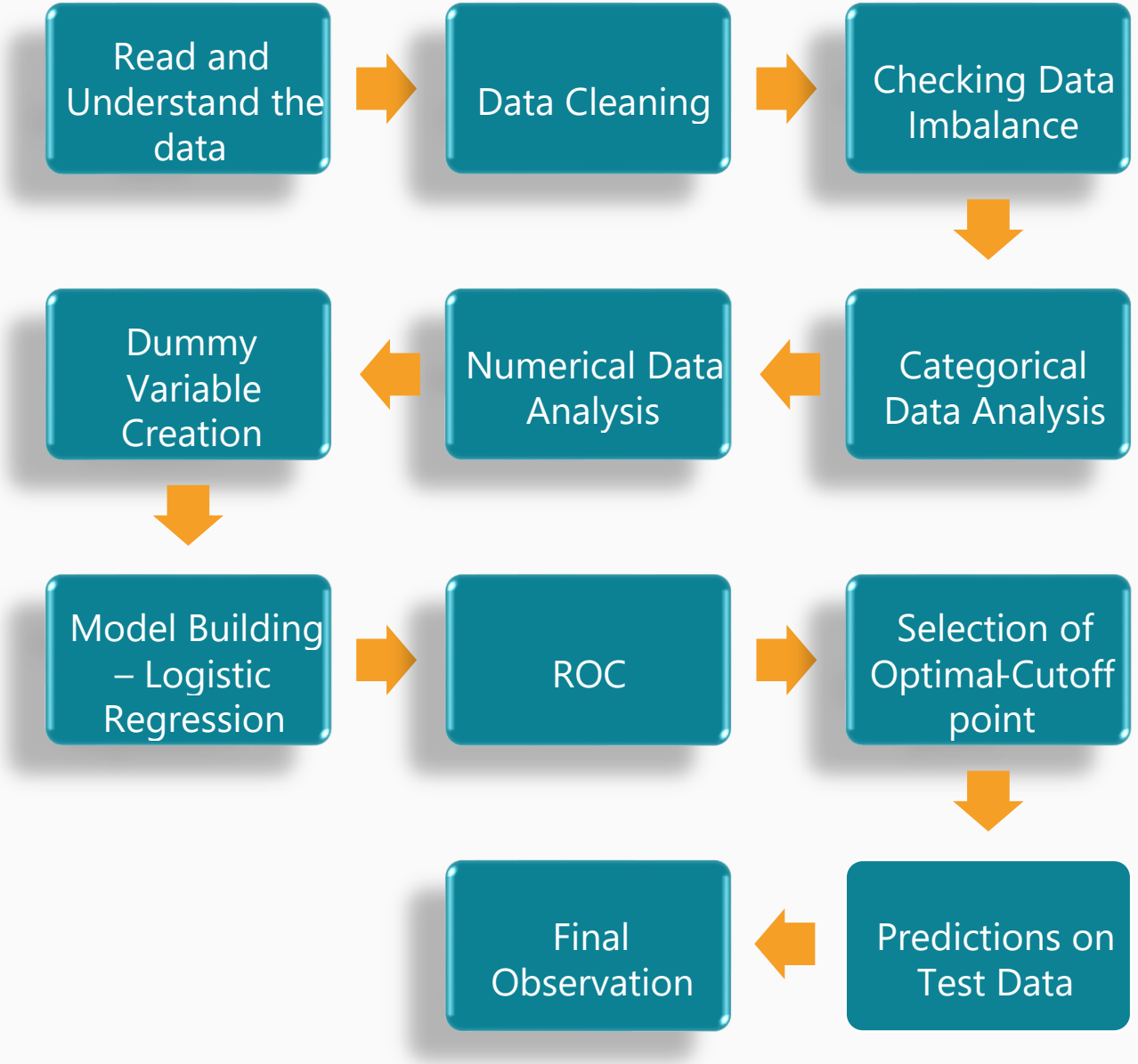
Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If 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.

# Business Objective

THERE ARE LOTS OF LEADS GENERATED AT THE THE CEO , HAS GIVEN A BALLPARK OF THE TARGET INITIAL STAGE BUT ONLY FEW OF THEM COME OUT LEAD CONVERSION RATE TO BE AROUND 80%. AS PAYING CUSTOMERS. CURRENT LEAD CONVERSION IS AROUND 30%.

OUR GOAL IS TO INCREASE LEAD CONVERSION RATE FROM AROUND 30% TO 80%. AND BUILDING A RIGHT MODEL TO IDENTIFY AND CLASSIFY THE MOST POTENTIAL LEADS.

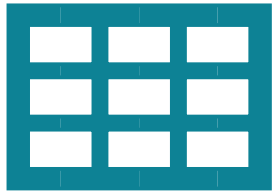
# Solution Methodology



# Understanding Data

## Reading Data :

Reading Dataset Leads.csv.



## Data Cleaning :

Checking Shape, Datatypes and Statistical Summary for the data frame.

Checking the presence of Duplicate values in the Data frame.

Dropped Unnecessary columns from the Data frame.

Data frame has ‘Select’ value for null values by default. So converted these ‘select’ values with null values.

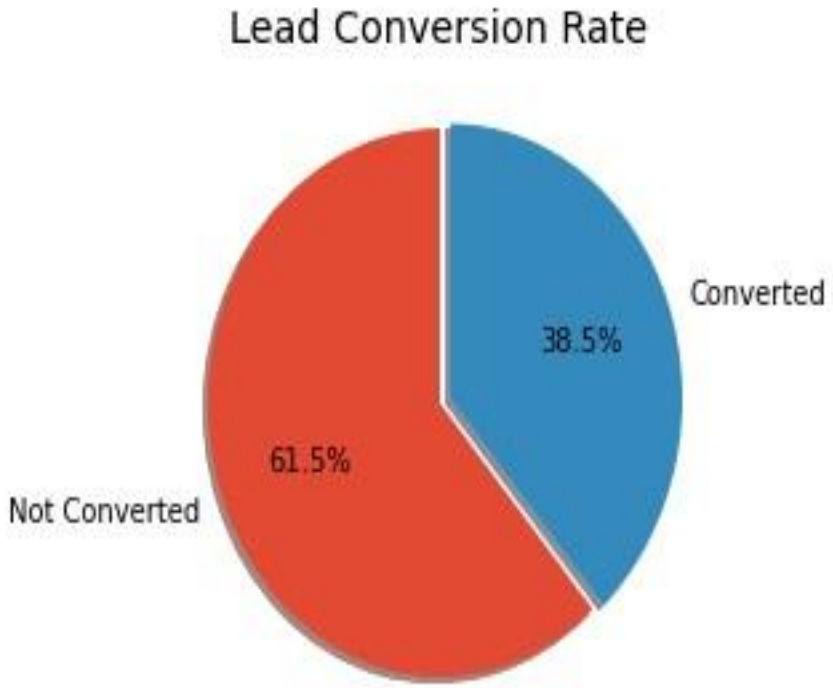
Checking the percentage of null values in the Data frame.

Dropped columns having null values > 45%.

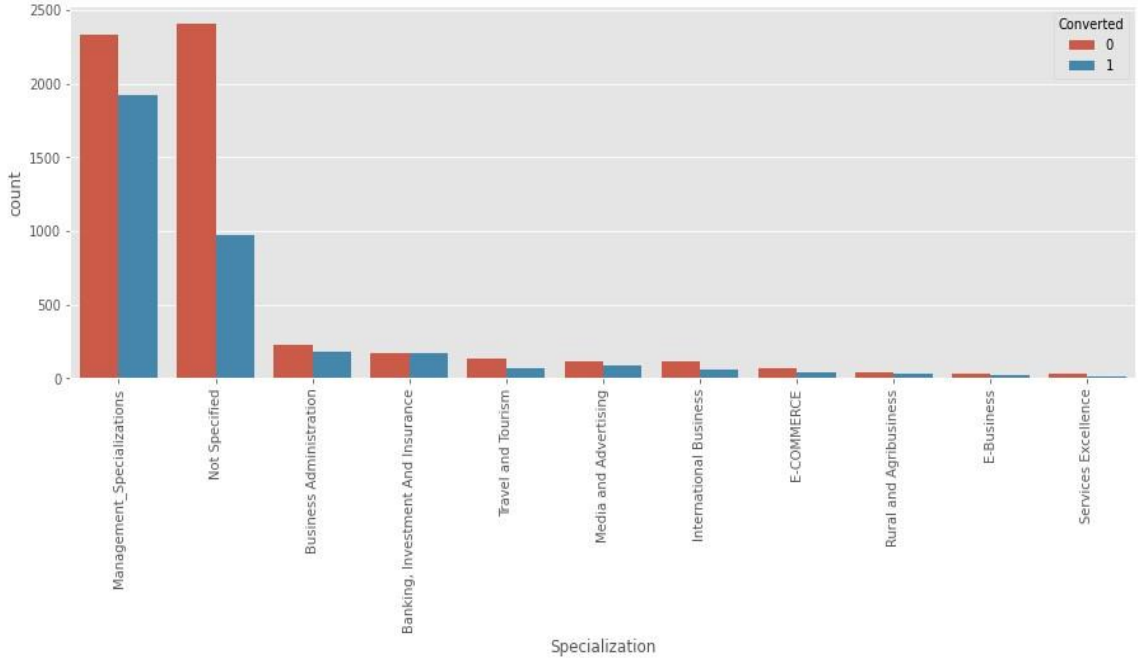
# Checking Data Imbalance

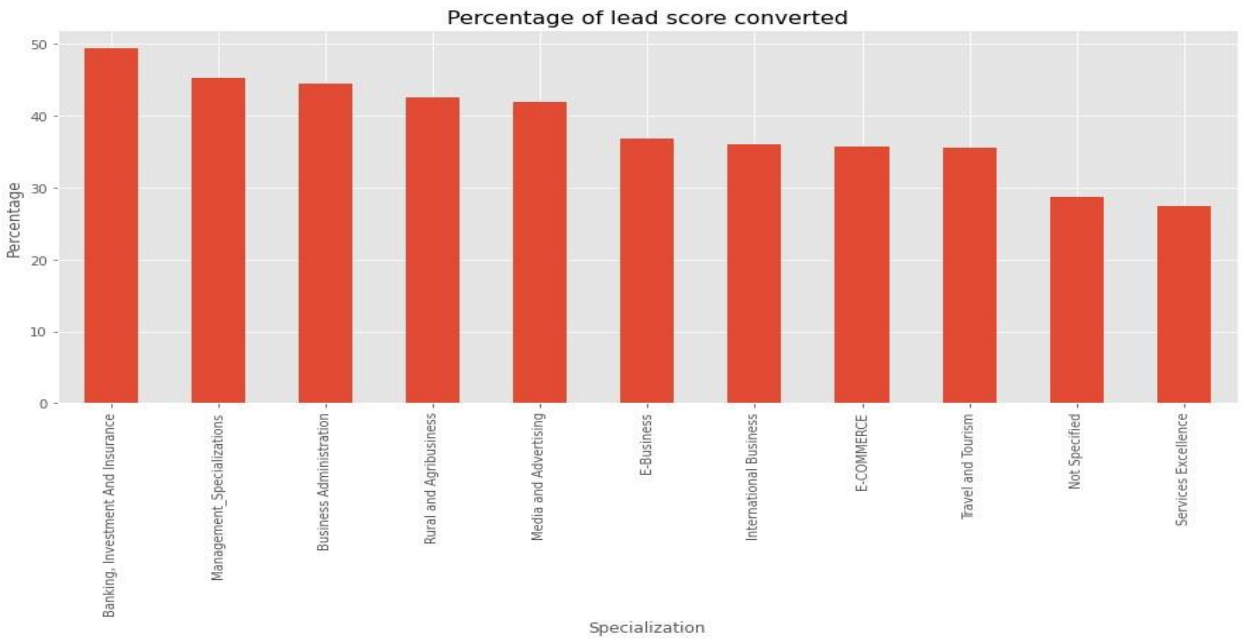
The Target class is not highly imbalanced

- 61.5% of observations as "0" - labeled as non-Converted or unsuccessful leads
- 38.5% of observations as "1" - labeled as Converted or successful leads



# Categorical Data Analysis



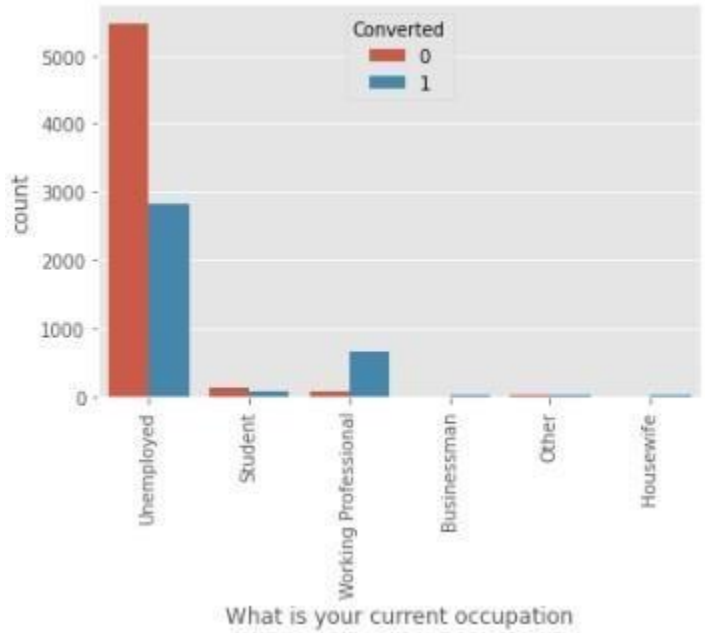



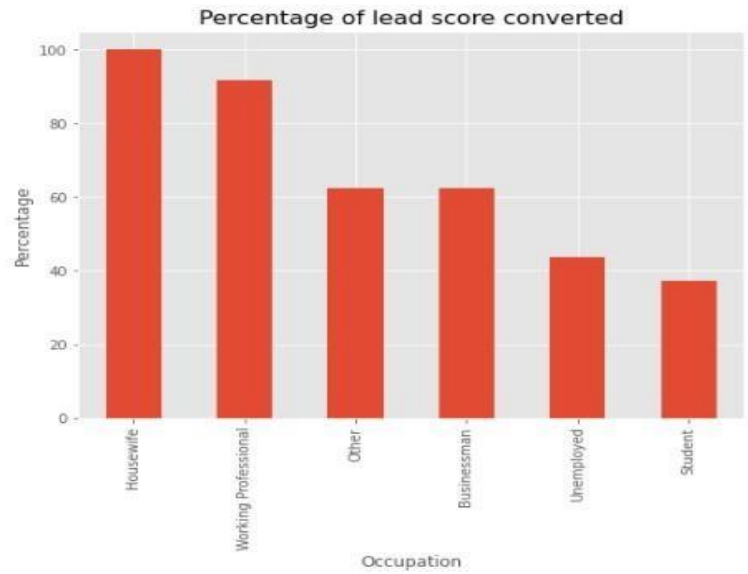

## Specialization :

- Maximum leads are generated from the Management Specialization .

- Though Specializations namely Banking, Investment and Insurance, Management Specialization, Business Administration, Rural and Agribusiness have higher conversion rate as compared to others.

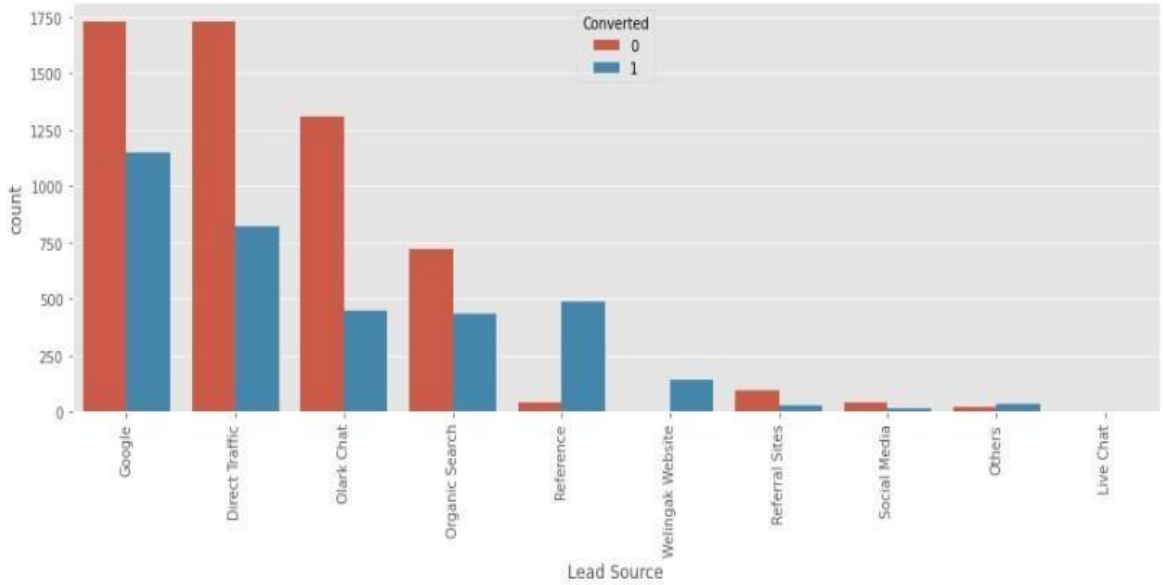
- Services Excellences have lowest conversion percentage.

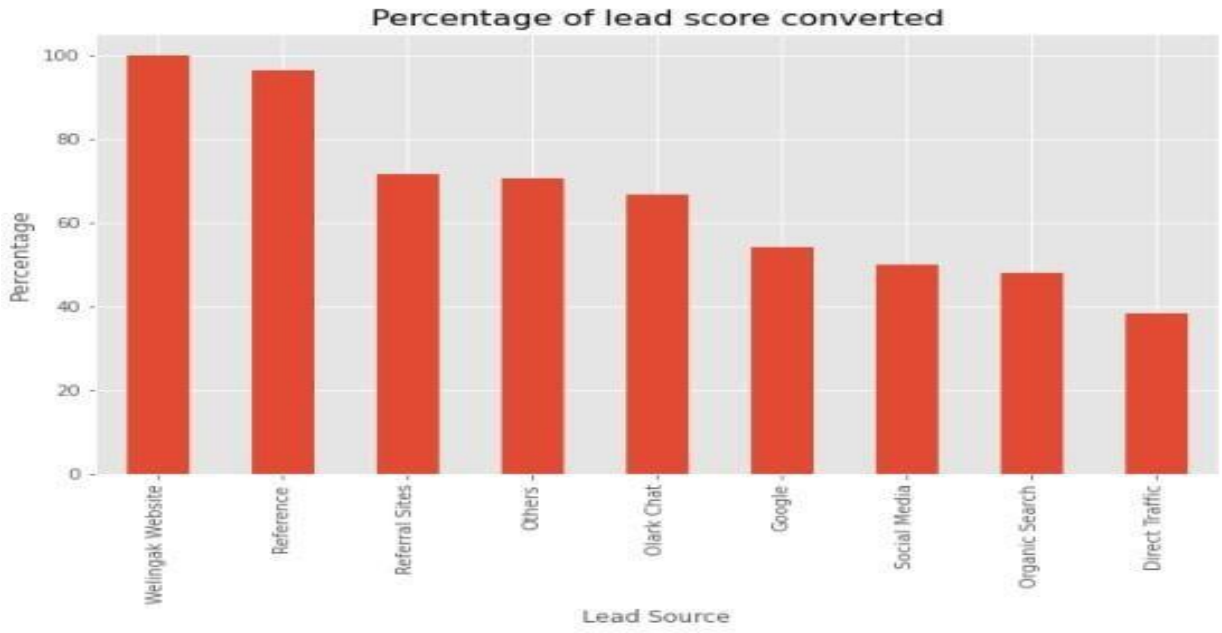



## Occupation :

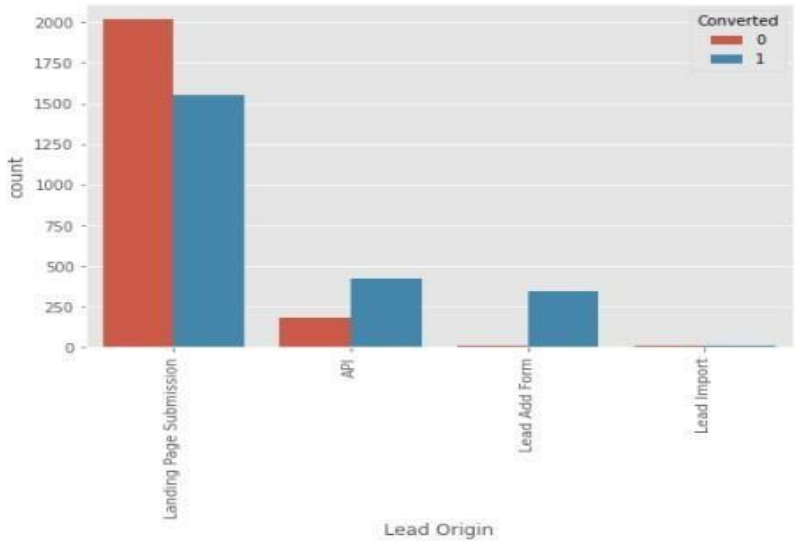
- Maximum leads are generated from the Unemployed People .
- Though **Housewife** have lowest frequency , but have 100% conversion rate
- Working professionals have higher conversion frequency.
- We should focus on Housewife and Working Professionals more as these two have high conversion rate.

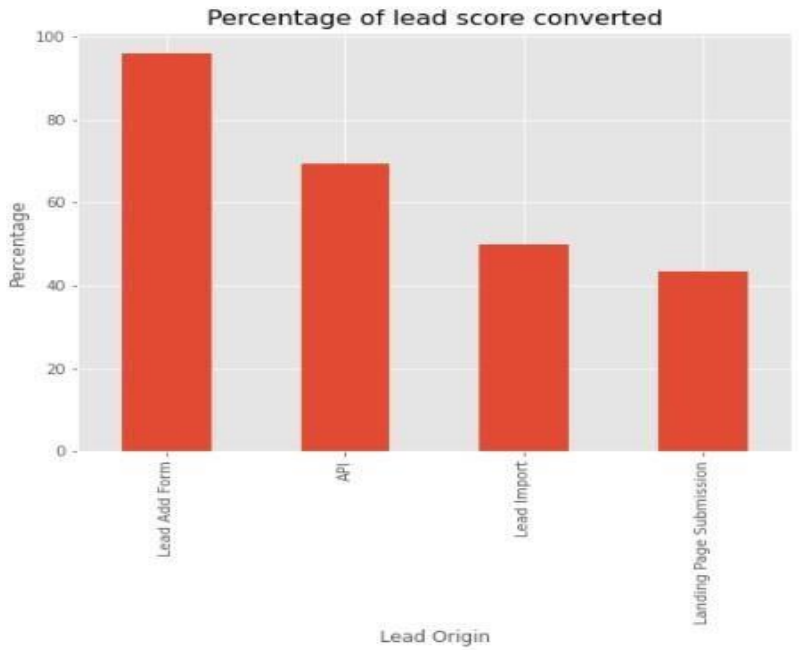



## Lead Source :

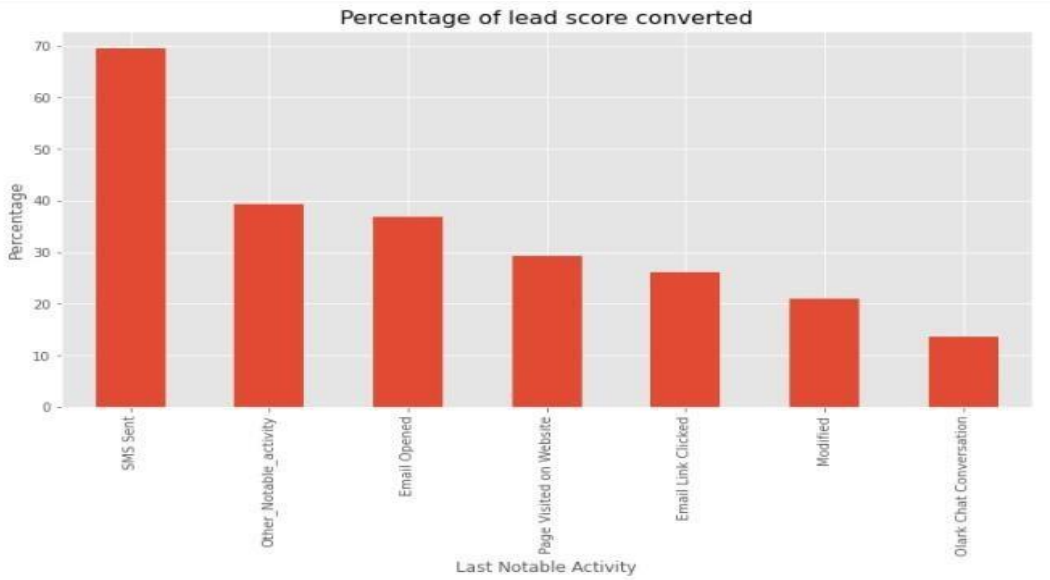
- Maximum number of leads are generated by Google and Direct traffic.
- Live Chat, Website , Reference leads with this lead sources have high conversion rate.
- Live chat have higher conversion rate though it have lowest frequency.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google lead sources and generate more lead sources from reference and welingak website.

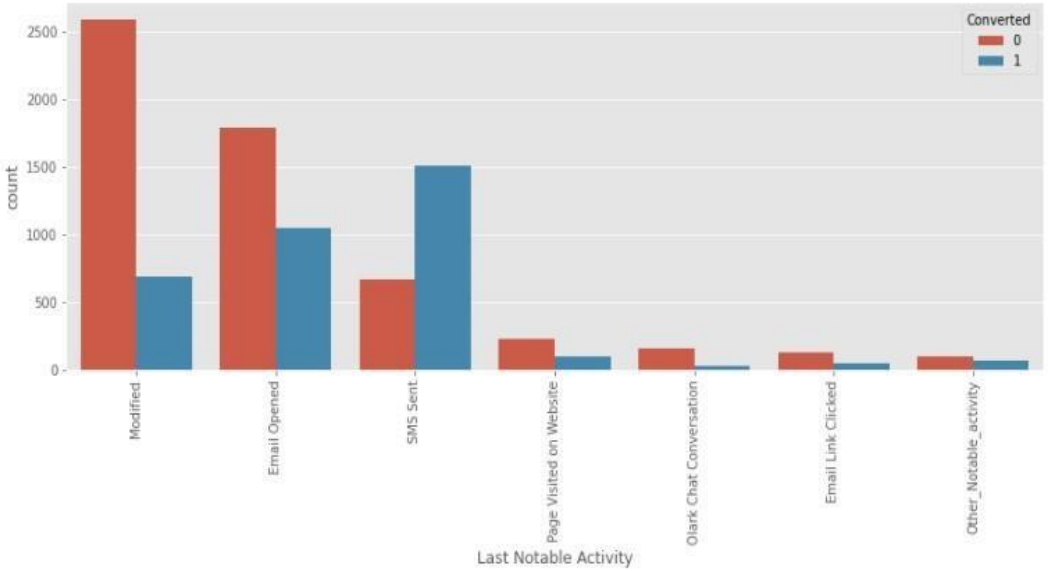



## Lead Origin:

- Maximum number of leads are generated by Google and Direct API and Landing Page Submission bring higher number of leads.
- Lead Add Form have very high conversion rate though total leads generated is lower.
- In order to improve overall lead conversion rate, we have to improve lead conversion of Land Page Submission and API. Also improve to generate mode leads from Lead Add Form.

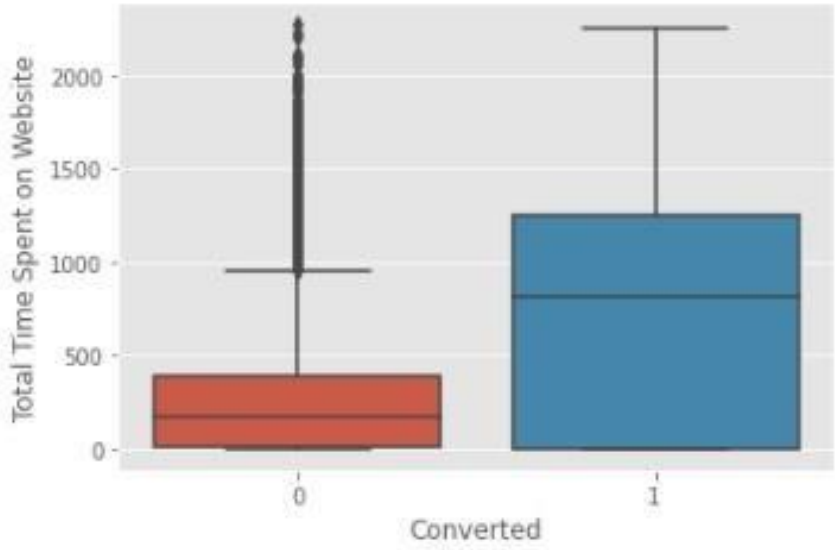


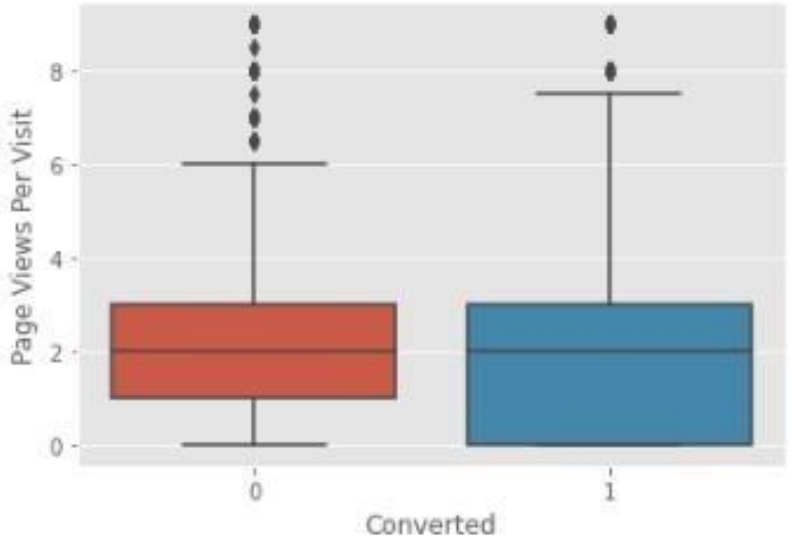




## Last Activity:

- Maximum leads generated by last activity as 'Email Opened' .
- Leads with last activity as 'SMS Sent' have the highest lead conversion rate .
- Leads are more likely respond to SMS .We should focus on this so as to increase conversion rates

# Numerical Data Analysis




# Page Views Per Visits:

- Spending more time on the website are more likely to be converted.
- Website should be made more engaging to make ls spend more time.

# Model Building

# Data Scaling

Features Converted

All Numeric Features Available in the Data Set

Method Used

**"Standardization "**
**Another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation**

Library Imported

.
"from sklearn.preprocessingmpo rt StandardScaler"

## Before RFE

```
[('Totalvisits',),
 ('Total Time Spent on Website',),
 ('Page Views Per Visit',),
 ('Lead Origin_Landing Page Submission',),
 ('Lead Origin_Lead Add Form',),
 ('Lead Origin_Lead Import',),
 ('What is your current occupation_Housewife',),
 ('What is your current occupation_Other',),
 ('What is your current occupation_Student',),
 ('What is your current occupation_Unemployed',),
 ('What is your current occupation_Working Professional',),
 ('City_Other Cities',),
 ('City_Other Cities of Maharashtra',),
 ('City_Other Metro Cities',),
 ('City_Thane & Outskirts',),
 ('City_Tier II Cities',),
 ('Specialization_Banking, Investment And Insurance',),
 ('Specialization_Business Administration',),
 ('Specialization_E-Business',),
 ('Specialization_E-COMMERCE',),
 ('Specialization_International Business',),
 ('Specialization_Management_Specializations',),
 ('Specialization_Media and Advertising',),
 ('Specialization_Rural and Agribusiness',),
 ('Specialization_Services Excellence',),
 ('Specialization_Travel and Tourism',),
 ('Lead Source_Direct Traffic',),
 ('Lead Source_Google',),
 ('Lead Source_Live Chat',),
 ('Lead Source_Olark Chat',),
 ('Lead Source_Organic Search',),
 ('Lead Source_Reference',),
 ('Lead Source_Referral Sites',),
 ('Lead Source_Social Media',),
 ('Lead Source_Welingak Website',),
 ('Last Activity_Converted to Lead',),
 ('Last Activity_Email Bounced',),
 ('Last Activity_Email Link Clicked',),
 ('Last Activity_Email Opened',),
 ('Last Activity_Form Submitted on Website',),
 ('Last Activity_Olark Chat Conversation',),
 ('Last Activity_Page Visited on Website',),
 ('Last Activity_SMS Sent',),
 ('Last Notable Activity_Email Link Clicked',),
 ('Last Notable Activity_Email Opened',),
 ('Last Notable Activity_Modified',),
 ('Last Notable Activity_Olark Chat Conversation',),
 ('Last Notable Activity_Page Visited on Website',),
 ('Last Notable Activity_SMS Sent',),
 ('Tags_Busy',),
 ('Tags_Closed by Horizzon',),
 ('Tags_Interested in other courses',),
 ('Tags_Lost to EINS',),
 ('Tags_Other_Tags',),
 ('Tags_Ringing',),
 ('Tags_Will revert after reading the email',)]
```

Total Features

RFE

## After RFE

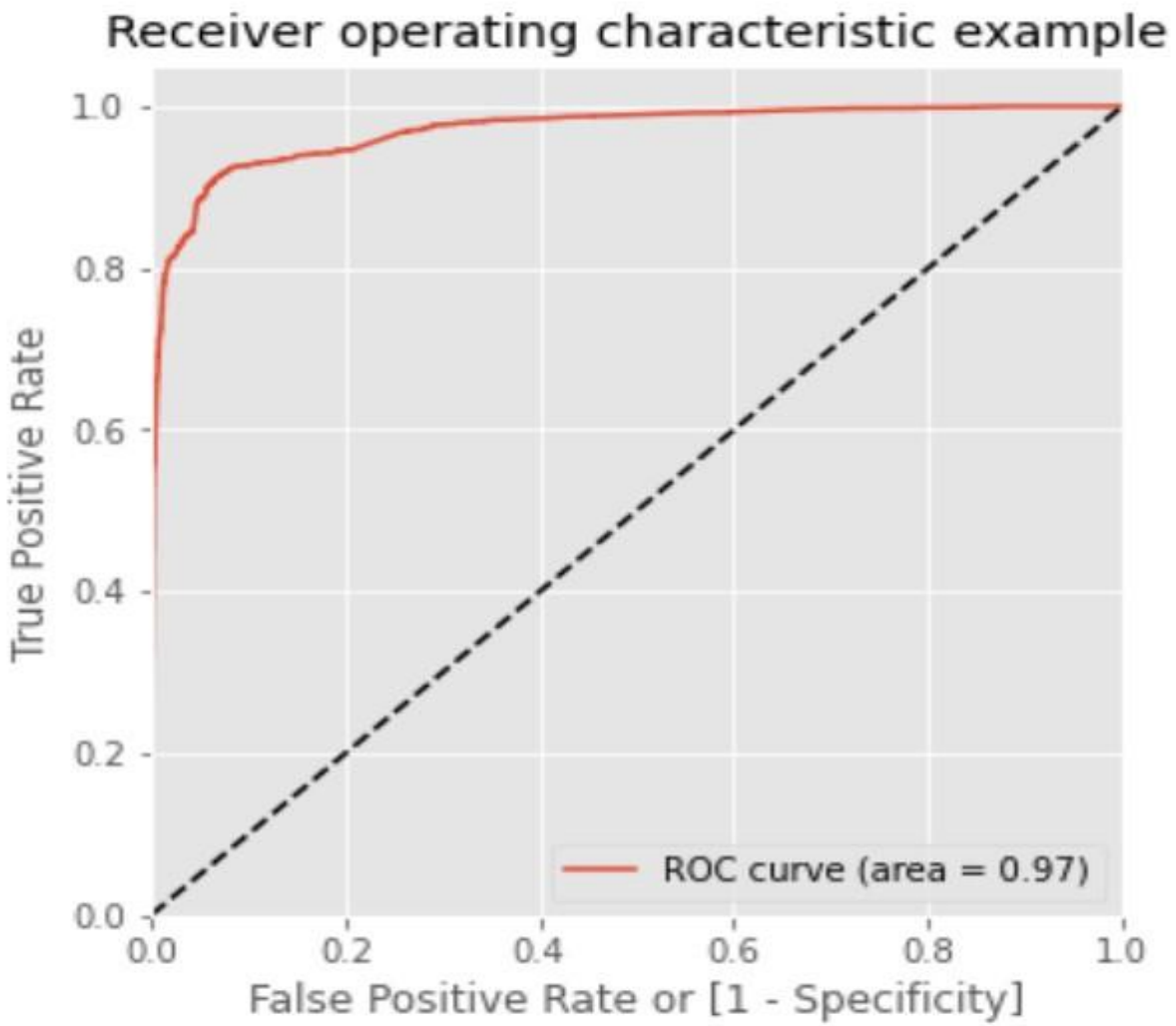
```
col = X_train.columns[rfe.support_]
list(col)
```

```
['Total Time Spent on Website',
 'Lead Origin_Lead Add Form',
 'Lead Source_Direct Traffic',
 'Lead Source_Referral Sites',
 'Lead Source_Welingak Website',
 'Last Activity_SMS Sent',
 'Last Notable Activity_Modified',
 'Last Notable Activity_Olark Chat Conversation',
 'Last Notable Activity_SMS Sent',
 'Tags_Closed by Horizzon',
 'Tags_Interested in other courses',
 'Tags_Lost to EINS',
 'Tags_Other_Tags',
 'Tags_Ringing',
 'Tags_Will revert after reading the email']
```

Out of Total Features , Top 15 Features Selected By RFE

# Feature Selection Using RFE

# ROC




## Why ROC?

The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

## At Cut off ( 50% )

| Accuracy | Sensitivity | Specificity |
|---|---|---|
| 92% | 88% | 95% |

## Area Under Curve (AUC)

0.97%

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

## Optimal Cutoff

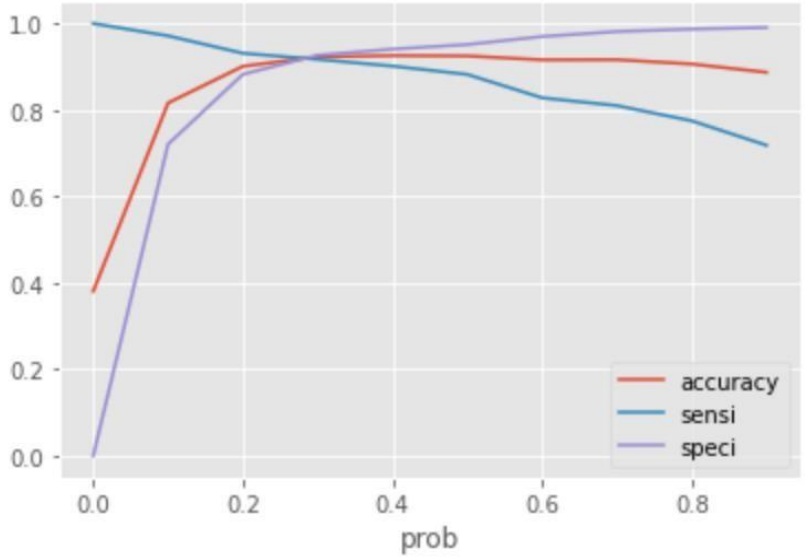



Fig.1: The graph shows a trade-off between Accuracy,Sensitivity,Specificity

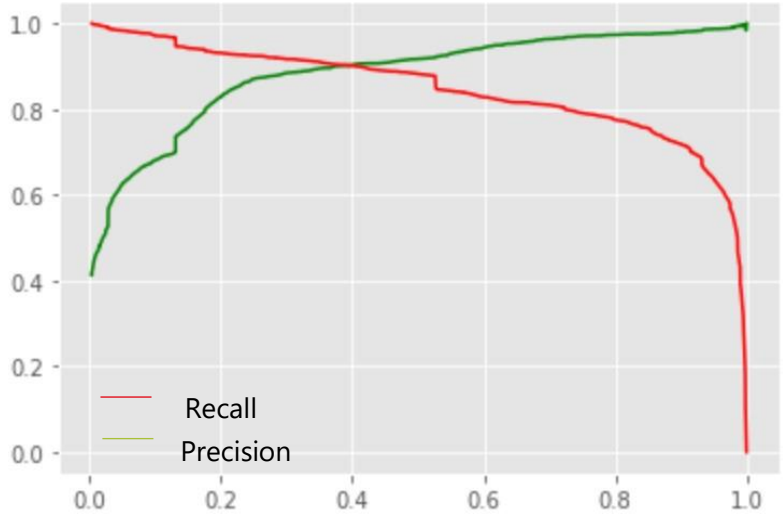



Fig.2: The graph shows a tradeoff between Precision and Recall

## Optimal Cut-off

0.30%

As per business requirement, we have chosen 0.30 as a Cut-Off value, which gives better results for both accuracy and precision

## At Optimal Cut-off (0.3)

**Accuracy**

92%

**Sensitivity**

91.6%

**Specificity**

92.6%

## Precision /Recall Rate at cut off 0.3

**Precision**

88.4%

**Recall**

91.6%

# Prediction Results/Observation

**At Optimal Cut-off (0.3)**

| | Accuracy | Confusion Matrix | Final Observation |
|---|---|---|---|
| **Train Set** | 92.2% | array([[3597, 285], [ 198, 2187]]) | Sensitivity 91.7% \| Specificity 93% \| Precision 88% \| Recall 92% |
| **Test Set** | 92.7% | array([[1563, 113], [ 81, 929]]) | 91.8% \| 93% \| 89% \| 92% |

# Lead Score and Conversion Rates

## Steps taken to assign a lead score variable for all customers.

### 1 Train the data with the model.

- Run the model on the entire Leads dataset
- Do not divide into Test and Train and run the obtained LR model on the entire data frame

### 2 Predict the Conversion Probability using Cutoff

- Predict the Conversion probability for all the customers using the cutoff value = 0.30.
- Create a new data frame and store the Conversion_Probability and actual converted values in this.

### 3 Adding Lead Score for all variables.

- Create a new column called Lead Score.
- Convert the probability score into Lead Score by multiplying by 100 and store it in this column.

### 4 Calculate the conversion rate.

- Once we obtain the complete model result on the data, we filter only the leads as predicted by the model.
- Calculate the Conversion Rate using this filtered result.

## Short Notes

**Conversion Rate is the number of customers who are converted to leads and interested in the course.**

**Before model building the Conversion Rate was found to be 38%**

**After model building, the conversion rate is increased to 88%**

**Hence we can conclude that our final model has served to the business purpose.**

# Hot Leads

HOT LEADS

1. *Hot leads are people who have a high probability to be converted as a Lead and thus needs to be identified. They have a higher conversion rate.*

2. *The leads whose lead score is greater than 30% are considered as potential leads. The conversion rate is around 88%. When we increase this threshold from 30% to 95%, we get Hot Leads.*

3. *Conversion Rate for hot leads is increases from 88% to 98%. This means they have a 98% probability of getting converted to a lead.*

4. *Focusing on Hot Leads will increase the chances of obtaining more value to the business as the number of people we contact are less but the conversion rate is high.*

# Conclusion

The Model seems to predict the Conversion Rate very well and we should be able to give the CEO

confidence in making good calls based on this model

- The model is prepared for prediction of the conversion of the leads. The probability values are generated by the model
- The cutoff decided for the model is 0.3, All leads whose probability is generated above this threshold value can be classified as Hot Lead

*It's good to collect data often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.*

*Along with phone calls, it's good to mail the leads also to keep them reminding as email is as powerful as cold calling.*


Thank You