

Summary Report

Reading the data and Null value analysis

Below columns have been dropped.

1. **Do Not Call:** Only 2 records opted for Yes and rest opted No for column 'Do not call'.
2. **Country:** More the 96% values are 'India'.

Below Columns were dropped Since these columns had only one value or large percentage of values are same

1. **Search**
2. **Magazine**
3. **Newspaper Article**
4. **X Education Forums**
5. **Newspaper**
6. **Digital Advertisement**
7. **Through Recommendations**
8. **Receive More Updates About Our Courses**
9. **Update me on Supply Chain Content**
10. **Get updates on DM Content**
11. **I agree to pay the amount through cheque**

EDA and Visualization

Below columns show some insights into Lead Conversion

Lead Origin: Highest conversion in 'Landing Page Submission' followed by 'API'

Lead Source: Highest conversion from 'Google' followed by 'Direct Traffic'

Last Activity: Conversion is highest amongst students who has 'SMS sent'

Country: Highest conversion has happened in India

Specialization: Highest conversion is seen for 'Not selected', this could mean these customers were not employed previously.

What is the current occupation: This data seems shows the maximum number of conversion happened in 'Unemployed' category

Tags: Highest conversion is seen in for 'Will revert after reading the email' option.

City: Mumbai has the highest conversion rate.

Last Notable Activity: Highest conversion is seen for SMS sent and Email Opened activities.

Correlations:

The Target column 'Converted' is highly correlated to the below features.

'Total time spent on Website'

'Lead Origin_Lead Add Form'

'Lead Source_Reference'
'Lead Source_Welingak Website'
'Last Activity_SMSSent'
'What is you current occupation_Working Professional'
'Lead Profile_Potential Lead'

Data Preparation

One hot encoding was done using pd.get_dummies on the categorical values.

Train Test Split:

The data was split into 70% Training data and 30% Test data using the train_test_split method from sklearn

Scaling:

Scaling of the continuous variables was done using MinMaxScaler method from sklearn

Correlated features:

A correlation matrix was created from the scaled features, the below features were highly correlated to each other

- 1.'Lead Source_Other' to 'Lead Origin_Lead Add Form'
- 2.'Lead Source_Reference' to 'Lead Origin_Lead Add Form'
- 3.'City_Not_selected' to 'Specilization_Not Selected'

Logistic Model building

GLM modelling from statsmodels was used to create the logistic model. This model provides summary of the model

Feature selection using RFE

15 features were selected.

```
1 : Do Not Email
2 : TotalVisits
3 : Total Time Spent on Website
4 : Lead Origin_Lead Add Form
5 : Lead Source_Olark Chat
6 : Lead Source_Welingak Website
7 : Last Activity_Not Known
8 : Last Activity_Olark Chat Conversation
9 : Last Activity_SMS Sent
10 : What is your current occupation_Working Professional
11 : Lead Profile_Not Selected
12 : Lead Profile_Other Leads
13 : Lead Profile_Student of SomeSchool
```

VIF was performed to find multi-collinearity between the features. Also, the p-values of the features were analysed over multiple iterations of model training and the below features were dropped.

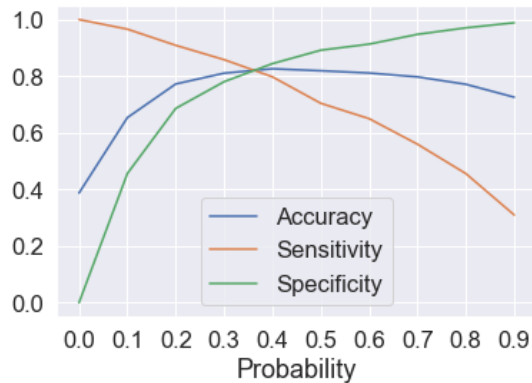
1: 'What is your current occupation_Housewife'

2: 'Lead Profile_Lateral Student'

Evaluating the model:

For the created model the below metrics were observed.

- 1. Accuracy: 0.82**
- 2. Sensitivity: 0.82**
- 3. Specificity: 0.82**
- 4. False positive rate: 0.18**
- 5. Positive Predictive Value : 0.82**
- 6. Negative Predictive Value:0.82**
- 7. Precision: 0.75**
- 8. Recall: 0.82**



From the curve above, 0.36 is the optimum point to take it as a cutoff probability.

A probability cut-off of 0.36 was based on the above curve for different probabilities.

Model Evaluation on the Test data:

- 1. Accuracy: 0.81**
- 2. Sensitivity: 0.8**
- 3. Specificity: 0.82**
- 4. Precision: 0.73**
- 5. Recall: 0.80**
- 6. F1 score: 0.77**

The co-efficients of the features after building the model is as below:

Total Time Spent on Website	4.609040
TotalVisits	4.489624
Lead Source_Welingak Website	3.526747
Lead Origin_Lead Add Form	3.466182
What is your current occupation_Working Professional	2.269519
Lead Source_Olark Chat	1.467893
Last Activity_SMS Sent	1.360855
const	-1.057006
Lead Profile_Other Leads	-1.244122
Last Activity_Not Known	-1.334708
Last Activity_Olark Chat Conversation	-1.406990
Do Not Email	-1.457342
Lead Profile_Not Selected	-1.890194
Lead Profile_Student of SomeSchool	-4.024844