------------------------------------Summary report---

Goal

To increase the conversion rate of leads from current 30% to 80% using data science techniques/analysis. To achieve 80% target, score between 0 and 100 to be assigned to each of the leads which can be used by the company to target potential leads. A higher score (>50%) would mean that the lead is hot and most likely it will convert whereas a lower score would mean that most likely lead will not convert.

Materials and Methods

Data provided by company has been used that includes details of 9000 site vistors gathered from website analytics. Collected data comprises of the activities performed by the visitors i.e. when people land on the website, what courses they browse, when they watch some videos, and/or the inputs provided while filling up the form for the course. When visitors fill up a form providing their email address or phone number, they are classified to be a lead, so that the call centre folks can approach them through email or phone. Moreover, the company also gets leads through referrals.

Followed these steps to get meaningful insights

1. Reading and Understanding data: 36 data points were provided for every lead.
2. Data Cleaning and Data Manipulation
    a. Check and handle duplicate data.
    b. Check and handle N/A and missing values.
    c. Drop columns, if it contains large number of missing values and not useful for further analysis.
    d. Imputation of values, if needed.
    e. Check for outliers and outlier treatment.

    Based on the above, Single value features like “Update me on supply chain content”, “Magazine”, “Receive more updates about our courses”, “I agree to pay the amount through cheque”, “Get updates on DM Content” , Do Not Call”, “Search”, “Newspaper” have been dropped.

3. EDA
    a. Univariate data analysis: value count, distribution of variables etc.
    b. Bivariate data analysis: correlation between variables.

4. Model Building
    a. Feature scaling and creation of dummy variables.
    b. Classification techniques: Logistic regression used for the model and prediction.
    c. Validation of the model: Using confusion metrics and ROC curve, Precision and recall.

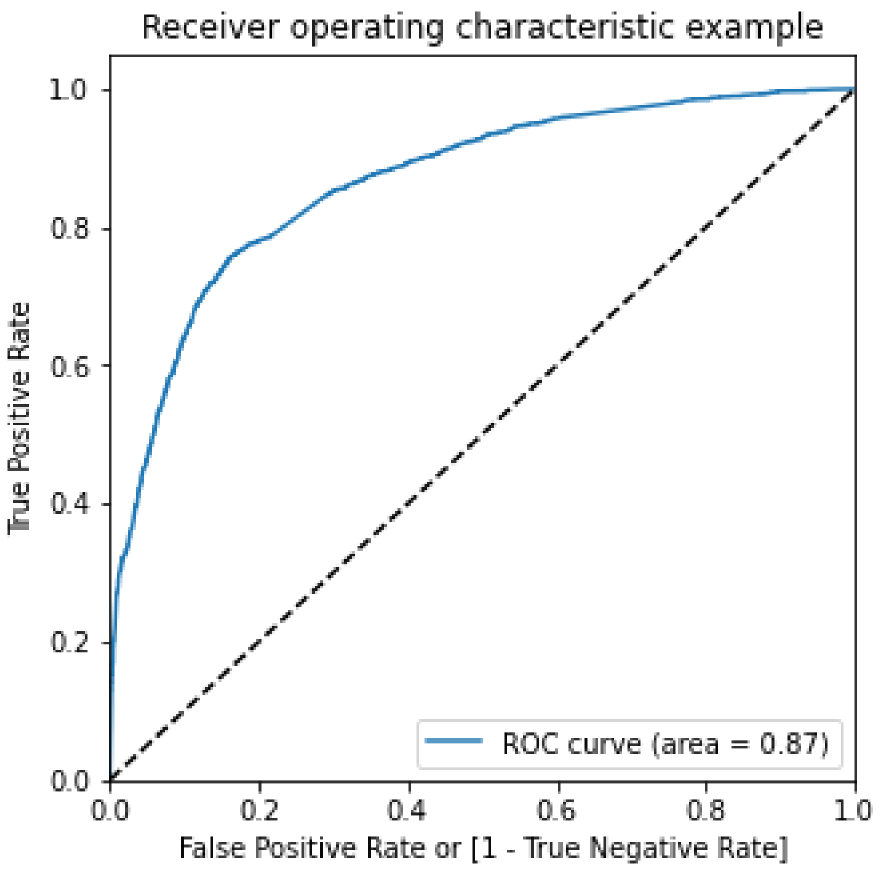
Model analysis

We have obtained good ROC curve for this Model

It shows the trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test

The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Conclusion/Recommendation

The important features responsible for good conversion rate or the one's which contribute more towards the probability of a lead getting converted are:

a) Total Time Spent : Recommendation is to add more features/knowledge on site that visitors spent more time having correct/engaging information.
b) Lead_Origin_Lead Add Form : The form filling should be reachable with 1 click on every page so that more visitors fill the form. Also the current form could be shortened so as to have more complete fill rate.
c) LeadSource_Welingak Website : The commercials should be looked at with Welingak so that we get more leads from this site.

As overall accuracy of model is 80 percent, sensitivity is 66 percent and specificity is 89 percent, model seems to be good fit predict the lead probability to conversion. The model, based on accuracy, achieves the business goals as it meets the target of 80% conversion rate.

As we get more data, we can train the model again, and could get improved model and corresponding prediction of leads conversion.

Also we can add more features by collecting the data point for each lead so as to improve the accuracy.