Summary Report

Problem statement :

To identify the set of leads of X Education so that the lead conversion rate rises and the sales team of the company can focus more on communication with the potential leads (hot leads) rather than making calls to every customer.

The company requires us to build a model wherein we need 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 conversion chance and the customers with lower lead score have a lower conversion chance. We also have a ballpark of the target lead conversion rate to be around 80%

Analysis Approach :

- **Data Understanding:** This is the initial step to start our entire process as we need to understand our data precisely so as to go ahead with the cleaning and analysis parts. Here, we are performing the following steps to understand our data:

  1. Loading the dataset

  2. Checking the shape (number of rows and columns) of the data.

  3. Checking the columns and the statistical summary of the numerical columns

  4. Checking the info to see the types of the feature variables and the null values present. If we find any missing values we will clean the data in our later stage.

  We found quite a few categorical variables present in this dataset for which we will need to create dummy variables. Also, there are a lot of null values present as well, so we will need to treat them accordingly.

- **Data Cleaning and Preparation:** As we have clearly understood our data and what needs to be done to get a cleaner version so that our model can be created as accurately as possible we will move ahead with cleaning the data set. Here, we are performing the following steps to clean and prepare our data:

  1. Checking the number of missing values and outliers in each column.

  2. As we check the number of missing values, we find some columns having missing values more than 50%, as a result we drop them for cleaner data and for easier evaluation.

3. We also drop some unnecessary columns which are not relevant for our analysis and now we have our final clean dataset on which we can build our model

- **Prepare data for modelling :**

1. Creating dummy variables for categorical columns.
2. Data split into train and test split.
3. Scaling the data using MinMaxScaler.

- **Model Building :**

1. Creating the model with all the features.
2. Removing the insignificant features using RFE and manual approach(VIF and p-values).
3. Refit the model and checking again the VIF and p-values until we get satisfactory values.

- **Model Evaluation :**

1. Predict the customer conversion with optimal probability cutoff.
2. Measuring accuracy, sensitivity and specificity .
3. Measuring the Gini of the model.

- **Making Predictions on the Test Set :**

1. Test the model using test data set with the final created model in train set.
2. Measure accuracy, sensitivity and specificity in the test set.
3. Comparing the above measures with the train set.

Model Outcome :

- **Features of the final model :**

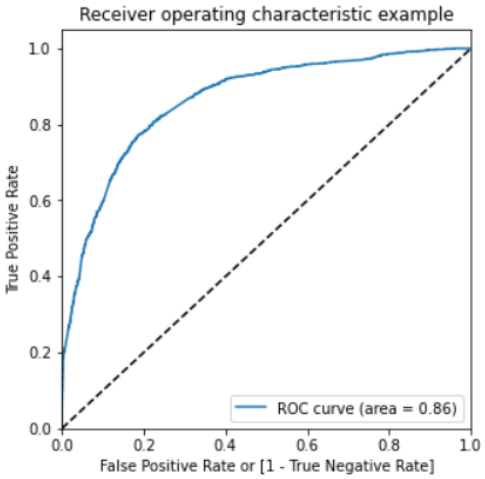
	Features	VIF
0	const	8.73
3	Page Views Per Visit	2.46
1	TotalVisits	2.14
4	Lead Origin_Lead Add Form	1.85
5	Lead Source_Olark Chat	1.74
2	Total Time Spent on Website	1.30
6	Lead Source_Welingak Website	1.28
10	Last Activity_Olark Chat Conversation	1.14
8	Last Activity_Converted to Lead	1.11
11	Last Activity_SMS Sent	1.11
12	What is your current occupation_Working Profes...	1.08
7	Do Not Email_Yes	1.02
9	Last Activity_Had a Phone Conversation	1.01

** const is not considered, so our final features are the ones except “const”.

The **coefficients** of the features to ensure positive and negative impact.

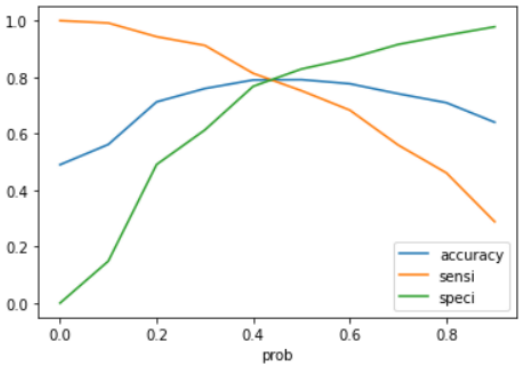
	coef
const	-2.0851
TotalVisits	1.7978
Total Time Spent on Website	4.3571
Page Views Per Visit	-1.0628
Lead Origin_Lead Add Form	3.6233
Lead Source_Olark Chat	1.5637
Lead Source_Welingak Website	2.5528
Do Not Email_Yes	-1.4353
Last Activity_Converted to Lead	-0.9711
Last Activity_Had a Phone Conversation	1.6958
Last Activity_Olark Chat Conversation	-1.2919
Last Activity_SMS Sent	1.1159
What is your current occupation_Working Professional	2.5789

- **Gini of the model : 0.86**



Here we see that the **Area under the ROC Curve is 0.86** which in our opinion is quite good.

Optimum probability cut off: **0.42**



Accuracy of trained model : 79%

We will compare this value with our test data

So, now we follow the same procedure for our test data and our final test data set is:

	Converted	Conversion_Prob	final_predicted
0	0	0.066097	0
1	0	0.047943	0
2	1	0.694944	1
3	0	0.058205	0
4	1	0.644389	1

Accuracy of test model : 78%

- ## Conclusion:

Comparing the accuracy of the train and test we can come to the conclusion that the model has good accuracy, sensitivity and specificity. Overall, the model performs well inthe test set, what it had learnt from the train set.

- ## Business Recommendations:

1. Analyze the total number of visits, if the number is on the higher side the leads can be potential
2. Total time spent by a candidate on the website can be taken as a measure to understand potential leads
3. Customers opted for 'Do not email' option are very less likely to be converted to leads
4. Last activity of the customers is any of 'Olark chat conversation' are very less likely to be converted to leads

- ## Learnings Gathered:

1. **Data preparation for modelling**
   a) It is important to treat missing values and also get rid of the outliers present in the data.
   b) If there is huge data imbalance in the features, then it is better to either drop that particular feature or remove the imbalance by merging the imbalanced values to other values.
   c) All the features should be in the same scale.

2. **Model building**
   a) There shouldn't be any multicollinearity between the variables.
   b) Find the optimal probability cut off to get a balance between Sensitivity andSpecificity with good Accuracy.
   c) The model should perform well in the test set in terms of Sensitivity,Specificity and Accuracy.