Summary

The analysis was done for X education with the target to identify the leads as hot leads and cold leads so as to improve their conversion rate to 80%. The dataset contained the basic info about these potential leads such as how they came to know about the course, how much time did they spend on the website, also how many time did they visited the website so on.

Following is the approach that we used in this problem :

1.Cleaning the data:

The dataset consist of various missing values and 'select' as one of the value which needed to be replaced as it was not providing any info. The "select' value was replaced by Nan value and the columns having missing or null values more than 45% were directly removed. The rest of the missing values were either imputed with the mode value or with "not specified" value. Columns having very minimal missing values were removed alongside the rows, reducing the total number of rows in the dataset. Other than this few column had dominating single values such as countries column had India as a value around 98% so this column was removed as it was causing imbalances, similarly few other columns were also removed. Column "Lead Origin" contains value as Google with "G" and "google" wioth "g" so the google was replaced by "Google".

2. EDA

In EDA part we analysed few numerical columns and found out they had few outliers, these outliers were treated by removing the top and bottom 1 percentiles of value.

3. Dummy variables:

All the categorical columns were assigned dummy variables and the original column was removed. For numerical values standard scaling was performed.

4. Train Test Split.

The dataset was divided into 80:20 ration for training and testing purpose.

5. Model Building:

Using the RFE model 15 features were identified. Later based on the p value and the Variance Inflation factor few columns were dropped. The value of p value was kept below <0.05 and VIF< 5.

6. Model Evaluation:

For predicting the model accuracy confusion matrix was used and the accuracy for training and testing set was found out to be nearly 92%

7. Recall :

The recall value for the model was approx 91% with 89% precision.

Top three variables in your model which contribute most towards the probability of a lead getting converted are :

- 'Total Time Spent on Website'
- 'Last_notable_activity_sms_sent'
- Tags_will revert after reading the email

Good strategy they should employ at this stage.

1. Make phone calls to the leads having lead score of greater than 60.

2. Make phone calls to the people spending more time on website and are currently working professionals.

3. Few Interns should be dedicated to make the website more engaging and user friendly

4. Leads with Last activity as SMS sent should be followed up regularly as they have high chance of being converted.