Summary Report

Problem Summary:

X educations sells online courses to industry professionals. Marketing of courses are done in multiple way like advertise in Search engine like google and in multiple website. Sales team got huge amount of lead but conversion rate is less as lead conversion rate is 30%.They want below

- To identify the Potential Lead.
- Build a model to assign lead score.
- Score the Lead by Conversion Probability Percentage
- Minimize marketing call to non-potential leads

Solution Overview:

Below are steps in solution flow we have followed to solve X education lead conversion problem.

- Data Analysis
- Data Cleaning
- Data Preparation
- Build Model
- Model Analysis and Selection of Variables
- Assign Lead Score
- Suggestion

Solution Steps:

Data Analysis:

- Lead data is available as our primary data source for this analysis.
- Checked the data and found that data frame has 9240 entries but few features are there with less number of rows.
- Checked the data type of the columns and converted Lead number to object.
- Found Duplicates in Lead source as google
- Many categorical columns have low variance and missing values.
- Few of the categorical column has rows with 'Select' value that may be the default option for the column.

Data Cleaning and Preparation:-

- Converted 'Select' to Others for Specialization and nan for Lead profile, City and How did you he ar about X education

- Dropped 13 categorical column for very low variance
- Dropped below columns as these has more than 30 % missing values
    - How did you hear about X Education
    - Lead Quality
    - Lead Profile
    - City
    - Asymmetrique Activity Index
    - Asymmetrique Profile Index
    - Asymmetrique Profile Score
    - Asymmetrique Activity Score
    - Tags
- Dropped rows which has more than or equal to 3 columns blank
- Arrived at a data frame of 9130 rows(98.8%)
- Imputed the categorical with highest occurring value and numerical with mean
- Capped the outliers for numerical variables
- Created dummy column for categorical variable and dropped low variance dummy column
- Split the data frame into train(70%) and test(30%)
- Scaled the data set using standard scaler

Model Building /Analysis and Selection:

- Built 3 models using logistic regression algorithm and selected the features using RFE
- Checked the VIF
- Using ROC curve chose cut-off and calculated specificity, sensitivity and accuracy
- Selected Model 1 as our preferred as we got 80-81% of accuracy on train and test data set in Model
- Below are the key metrics for model 1

```
Sensitivity:                     0.77
Specificity:                     0.83
False Positive Rate:             0.17
Positive Predictive Value:       0.74
Negative Predictive Value:       0.85
Precision Score:         0.74
Recall Score:                    0.77
```

Scoring and Recommendation:

Based on the model chosen, 40 % is cutoff for lead conversion

Lead conversion rate is peak when score is between 95-100, 85-90, 55-60 .

Top 3 favorable variables are

Lead Origin is Lead Add Form (positive)

Do Not Email 'Yes' (negatively)

Last Notable Activity is SMS Sent (positive)

Learnings:-

Below are the Key Learnings

- Treatment of 'Select' value for the columns:
    - Select value maybe the default value of the field.
    - Select could have been chosen because the person does not want to reveal the field value or else the field is not applicable or else Others option is not there. So conversion of Select to a comprehendible value is tricky.
    - Assuming that user did not want to share or field is not applicable, cases where others is present Select was replaced with nan. For other cases, Select was replaced with Others value.
- Missing value Treatment for the features:
    - It is very tricky as we don't know the business value of the columns in details.
    - Dropped the column which has more than 30 percent of missing values
- Imputation of missing values:
    - We have assumed that the missing values need to be replaced with the max occurrence value in the column for categorical data.
    - This assumption may actually deviate some of the results.