**Summary**

The model is to build a logistic regression model for X Education Company is to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. The data is provided with 9000 data points with various attributes. The data is evaluated and validate as to increase the potential leads and factors which increase the conversion rate.

The steps are as follows:

EDA (Exploratory Data Analysis):

* Initially we check the data which consists of null values and drop of attributes whose null value percentage is more than forty five.
* Imputing the null values with mode/ others/ not specified as these null values weighs importance in the column imputation.
* Grouping of records which has contributes similar meaning/ low percentage under one group for analysis.
* Check for the data imbalance on the target variable.
* Check the categorical and numerical variables are fit for evaluation part for further analysis
* Outliers treatment for the numerical variables and drop of records with low and high percentiles
* Finalizing the variables for model evaluation

Train-Test Split, Scaling:

* The data is divided in to Train and Test for model evaluation by splitting the data with 70 and 30 percentage.
* After splitting the data under Train and Test, min-max scaling is done on numerical variables and dummies are generated for categorical variables.

Model Building:

* By using sk learn and RFE feature selection is done.
* In RFE with selection features with 15 variables.
* Later the rest of the variables were removed manually depending on the VIF values and p-value.
* Finally the model is built with variables with low VIF and p-value.
* A confusion matrix and overall accuracy was checked which came out to be 87.55%.

Model Evaluation:

* Model evaluation is carried on both Sensitivity-specificity and Precision – Recall
* **Sensitivity – Specificity**

If we go with Sensitivity- Specificity Evaluation.

We will get:

* On **Training Data**
* The optimum cut off value was found using ROC curve. The area under ROC curve was 0.95.
* After Plotting we found that optimum cutoff was **0.30** which gave

Accuracy 87.55%

Sensitivity 81.00%

Specificity 94.61%.

* Prediction on **Test Data**
* We get

Accuracy 86.89%

Sensitivity 81.00%

Specificity 94.61%

* **Precision – Recall:**

If we go with Precision – Recall Evaluation

* On **Training Data**
* With the cutoff of 0.3 we get the Precision & Recall of 81.14% & 87.67% respectively.
* So to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of **0.4**.
* Prediction on **Test Data**
* With the cutoff of 0.3 we get the Precision & Recall of 90.35% & 86.23% respectively.
* So if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be **0.30**
* If we go with Precision – Recall Evaluation the optimal cut off value would be **0.40**

**CONCLUSION**

TOP VARIABLE CONTRIBUTING TO CONVERSION:

* Total Time Spent on Website:
  + Lead Origin:
    - Lead Add Form
  + Lead source:
    - Olark Chat
    - Welingak website
  + What is your current occupation
* Working Professional
  + Tags
* Will revert after reading the email
  + Last Notable Activity
* SMS Sent
* Others

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.