**SUMMARY REPORT**

Problem Statement:

• X Education sells online courses to industry professionals.

• Although X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%.

• To make this process more efficient, the company wishes to identify the potential leads, also known as ‘Hot Leads’.

• If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

• X Education wants to identify the most promising leads.

• The company wants to build a model to identify the hot leads.

• Improve the lead conversion rate from 30% to 80%

Solution Methodology

1. **Read and Understand Data**
2. **Data Cleaning:**

Data Cleaning is the very first step which needs to be performed on the dataset before building any model. The given dataset had a lot of NULL values which needed to be cleaned. We identified the NULL columns and dropped any Columns with more than 45% of Null Values. Also, the dataset had values ‘Select’ which is equivalent to NULL as the select value would be default if no value is selected by the user in the dropdown.

Next, we identified the highly skewed columns and dropped them. Once we have dropped the columns with high percentage of NULL values, the next step was to impute the missing values and identifying the outliers. We used various techniques such as Mean, Median Mode to impute the missing values.

1. **Exploratory Data Analysis:**

After the Data Cleaning was completed, we needed to do an Exploratory Data Analysis on the data. We found few outliers and performed the Univariate and Bivariate analysis on the dataset with respect to Target Variable.

1. **Dummy Variables:**

We created dummy variables for the categorical columns and dropped least relevant dummy for each category.

1. **Scaling:**

As the data was not scaled properly, we used the Standard Scaler to scale the data

1. **Train-Test Split:**

Train-Test split was done on the dataset in 70:30 ratio.

1. **Model Building:**

For Model building we used both Automated and Manual process. First we used RFE to provide with top 15 variables and then we manually dropped the insignificant variables based on the P-Value and VIF values.

1. **Model Evaluation:**

A confusion matrix was made. Later on the optimum cut-off value by using ROC curve was used. Area under the Curve came to be 87%.

1. **Prediction:**

We performed the Prediction on the test Dataset and found the below values:

* Accuracy: 80.0 %
* Sensitivity: 82.3%
* Specificity: 78.61%

1. **Precision-Recall:**

We used the precision-Recall method to recheck the cutoff of approx. 0.41.

1. **Conclusion:**

Training Set:

* Accuracy: 80.04%
* Sensitivity: 83.07%
* Specificity: 78.80%

Test Set:

* Accuracy: 80.08%
* Sensitivity: 82.38%
* Specificity: 78.61%

The ability of accurately predicting the conversion rate of any lead from the model is around 80%. The probability of predicting a promising lead from the model is around 82%.

The model seems to predict the Lead Conversion Rate very well. It should be able to give the CEO confidence in making good calls based on this model.

Variables that mattered most in the lead conversion are -

* Website\_Time\_Spent with an coefficient of 0.9222
* Lead Origin\_Lead Add Form with an coefficient of 3.8048
* Occupation\_Working Professional with an coefficient of 2.6938

As these parameters have positive coefficients, this will improve the lead score and in turn, these people will be ‘hot leads’ for X Education

Other Important Variables

* Do Not Email
* Last Activity\_Email Bounced
* Last Activity\_Olark Chat Conversation
* Last Notable Activity\_Email Link Clicked
* Last Notable Activity\_Email Opened
* Last Notable Activity\_Modified
* Last Notable Activity\_Olark Chat Conversation
* Last Notable Activity\_Page Visited on Website