**Summary**

**OBJECTIVE**:

• To build a logistic regression model for an education company named "X Education" by assigning a lead score between 0 and 100 for targeting particular leads to be converted or not.

• Adjusting the company's requirement for further changes in the future

**PROBLEM STATEMENT:**

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. X Education has appointed you to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you 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 lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

**METHODOLOGY:**

**Data Inspection and Cleaning:**

The data was partially clean except for a few null values and the option ‘select’ had to be replaced with a null value since it did not give us much information. And variables having missing values more than 45% were dropped. Few of categorical variables had vivid categories of data we have combined the less % of categories into ‘Others’ so as to not lose much data. This was later removed while making dummies.

**EDA and Data Preparation:**

EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and less outliers were found.

Dummy Variables: The dummy variables were created and later on the dummies with ‘others’ elements were removed. For numeric values we used the Standard Scalar.

Train-Test split: The split was done at 70% and 30% for train and test data respectively.

**Model Building:**

RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and P-value (The variables with VIF < 5 and p-value < 0.05 were kept).

**Model Evaluation:**

A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

**Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.335 with accuracy, sensitivity and specificity of around 80%.

**Lead Score:**

• The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.

• Higher the lead score, higher is the probability of a lead getting converted and vice versa,

• Since, we had used 0.335 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 33.5 or above will have a value of ‘1’ in the Final predicted column.

**Conclusion:**

• 15 features have been used by our model to successfully predict if a lead will get converted or not.

• The Coefficient (beta) values for each of these features from the model parameters are used to determine the order of importance of these features.

• Features with high positive beta values are the ones that contribute most towards the probability of a lead getting converted.

• Similarly, features with high negative beta values contribute the least.

**The magnitude and sign of the coefficients loaded in the logit function:**

**logit(p) = log(p/(1-p)) =**

**(3.42 \* Lead Origin\_Lead Add Form)**

**+ (2.84 \* what is your current occupation\_Working Professional)**

**+ (1.99 \* Lead Source\_Welingak Website)**

**+ (1.78 \* Last Activity\_SMS Sent)**

**+ (1.25 \* Last Activity\_Unsubscribed)**

**+ (1.09 \* Total Time Spent on Website)**

**+ (0.98 \* Lead Source\_Olark Chat)**

**+ (0.84 \* Last Activity\_Unreachable)**

**+ (0.66 \* Last Activity\_Email Opened)**

**- (0.25 \* Lead Origin\_Landing Page Submission)**

**-(0.87 \* Last Activity\_Olark Chat Conversation)**

**- (1.26 \* Do Not Email) -1.77**