Lead Scoring Case Study

To build a Logistic Regression Model to predict whether a lead for online courses for an education company named X Education would be successfully converted or not.

Business Objective:

The primary goal is to assist X Education in identifying the most promising leads, termed "Hot Leads," who exhibit a high likelihood of converting into paying customers. This involves constructing a logistic regression model that assigns a lead score ranging from 0 to 100 to each lead. These lead scores will aid the company in effectively targeting potential leads for conversion.

The objectives can be further categorized into the following sub-goals:

1. Developing a Logistic Regression Model: Build a logistic regression model to accurately predict the probability of lead conversion for each individual lead.
2. Calculating Lead Score: Multiply the lead conversion probability by a factor to determine the lead score for every lead. This lead score will serve as a quantifiable metric for evaluating lead potential.
3. Establishing Probability Threshold: Determine a threshold probability value above which a lead will be categorized as "converted." Conversely, leads falling below this threshold will be classified as "not converted."

By achieving these objectives, the aim is to enhance X Education's lead targeting strategy, allowing them to concentrate their efforts on leads with the highest potential for conversion.

Problem solving methodology

- The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented in a sequential flow as below.

1. Understanding the Data Set & Data Preparation
2. Applying Recursive feature elimination to identify the best performing subset of features for building the model.
3. Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model
4. Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.
5. Decide on the probability threshold value based on Optimal cutoff point and predict the dependent variable for the training data.
6. Use the model for prediction on the test dataset and perform model evaluation for the test set.

Data preparation and engineering

- The following data preparation processes were applied to make the data dependable so that it can provide significant business value by improving Decision Making Process:

1. Remove columns which has only one unique value- Deleting the following columns as they have only one unique value and hence cannot be responsible in predicting a successful lead case – ‘Magazine’, ‘Receive More Updates About Our Courses’ , ‘Update me on Supply Chain Content’ , ‘Update me on Supply Chain Content’ and ‘I agree to pay the amount through cheque’.
2. Removing rows where a particular column has high missing values- ‘**Lead Source’** is an important column for analysis. Hence all the rows that have null values for it were dropped.
3. Imputing NULL values with Median- The columns ‘TotalVisits’ and ‘Page Views Per Visit’ are continuous variables with outliers. Hence the null values for these columns were imputed with the column median values.
4. Imputing NULL values with Mode - The columns ‘Country’ is a categorical variable with some null values. Also majority of the records belong to the Country ‘India’. Thus imputed the null values for this with mode(most occuring value). Then binned rest of category into 'Outside India‘.
5. Handling 'Select' values in some columns: There are some columns in dataset which have a level/value called 'Select'. This might have happened because these fields in the website might be non mandatory fields with drop downs options for the customer to choose from. Amongst the dropdown values, the default option is probably 'Select' and since these aren't mandatory fields, many customer might have have chosen to leave it as the default value 'Select'. The Select values in columns were converted to Nulls.
6. Assigning a Unique Category to NULL/SELECT values- All the nulls in the columns were binned into a separate column ‘Unknown’. Instead of deleting columns with huge null value percentage(which results in loss of data), this strategy adds more information into the dataset and results in the change of variance. The Unknown levels for each of these columns will be finally dropped during dummy encoding.
7. Outlier Treatment : The outliers present in the columns 'TotalVisits‘ & 'Page Views Per Visit‘ were finally removed based on iterquatile range analysis.
8. Binary Encoding : Converting the following binary variables (Yes/No) to 0/1: **'Search'**,**'Do Not Email'**, **'Do Not Call'**, **'Newspaper Article'**, **'X Education Forums'**, **'Newspaper'**, **'Digital Advertisement'**,**'Through Recommendations'** and **'A free copy of**

Data preparation and engineering

9. Dummy Encoding - For the following categorical variables with multiple levels, dummy features (one-hot encoded) were created: 'Lead Quality','Asymmetrique Profile Index','Asymmetrique Activity Index','Tags','Lead Profile', 'Lead Origin','What is your current occupation', 'Specialization', 'City','Last Activity', 'Country‘ and 'Lead Source','Last Notable Activity‘
10. Test – Train split: The original dataframe was split into train and test dataset. The train dataset was used to train the model and test dataset was used to evaluate the model.
11. Feature scaling : Scaling helps in interpretation. It is important to have all variables(specially categorical ones which has values 0 and 1) on the same scale for the model to be easily interpretable. ‘Standardisation’ was used to scale the data for modelling. It basically brings all of the data into a standard normal distribution with mean at zero and standard deviation one.

Building the Model

- Generalized Linear Models from StatsModels is used to build the Logistic Regression model.
- The model is built initially with the 20 variables selected by RFE.
- Unwanted features are dropped serially after checking p values (< 0.5) and VIF (< 5) and model is built multiple times.
- The final model with 14 features, passes both the significance test and the multi-collinearity test.

	Features	VIF
1	Lead Origin_Lead Add Form	1.82
12	Tags_Will revert after reading the email	1.56
4	Last Activity_SMS Sent	1.46
5	Last Notable Activity_Modified	1.40
2	Lead Source_Direct Traffic	1.38
3	Lead Source_Welingak Website	1.34
10	Tags_Other_Tags	1.25
0	Total Time Spent on Website	1.22
7	Tags_Closed by Horizzon	1.21
11	Tags_Ringing	1.16
8	Tags_Interested in other courses	1.12
9	Tags_Lost to EINS	1.06
6	Last Notable Activity_Olark Chat Conversation	1.01

Predicting the conversion probability and predicted column

Creating a data frame with the actual Converted flag and the predicted probabilities.

Showing top 5 records of the data frame in the picture on the right.

	Converted	Converted_prob	Prospect ID
0	1	0.283149	9196
1	0	0.031440	4696
2	0	0.576636	3274
3	0	0.006433	2164
4	1	0.989105	1667

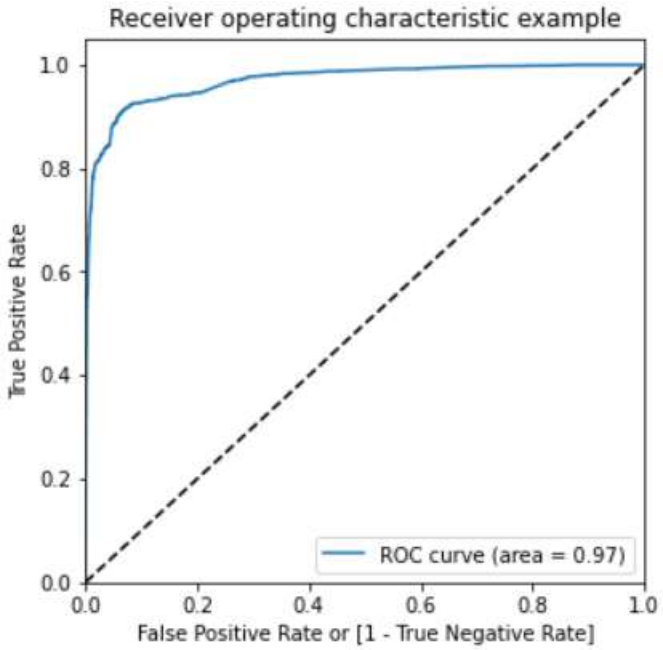
Creating new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0

Showing top 5 records of the data frame in the picture on the left.

	Converted	Converted_prob	Prospect ID	Predicted
0	1	0.283149	9196	0
1	0	0.031440	4696	0
2	0	0.576636	3274	1
3	0	0.006433	2164	0
4	1	0.989105	1667	1

Plotting ROC curve

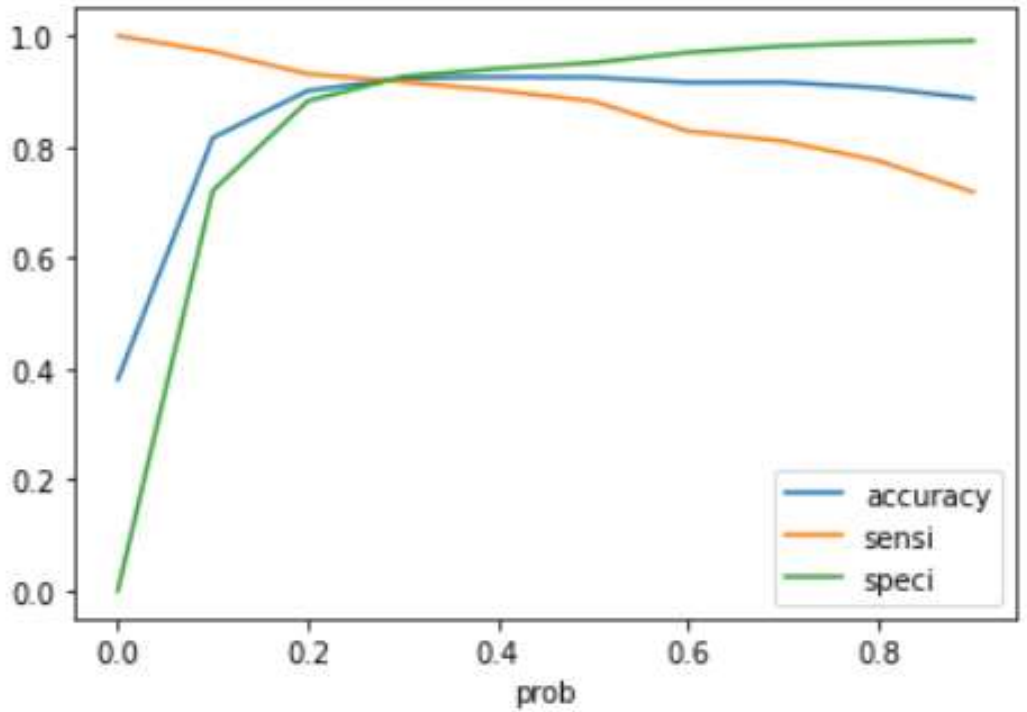
- Receiver Operating Characteristics (ROC) Curve : It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).



Evaluating the model on train dataset

We have the following values for the Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%



Making prediction on test data set

The final model on the train dataset is used to make predictions for the test dataset
The train data set was scaled using the scaler.transform function that was used to scale the train dataset.
The Predicted probabilities were added to the leads in the test data frame.
Using the probability threshold value of 0.33, the leads from the test dataset were predicted if they will convert or not.

	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	7681	0	0.024819	2	0
1	984	0	0.025692	3	0
2	8135	0	0.686054	69	1
3	6915	0	0.005880	1	0
4	2712	1	0.953208	95	1

The Conversion Matrix was calculated based on the Actual and Predicted 'Converted' columns.

After running the model on the Test Data these are the figures we obtain:

Accuracy : 92.78%
Sensitivity : 91.98%
Specificity : 93.26%