**Lead Scoring Case Study Summary**

**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.

Some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

To make this process more efficient, the company wishes to identify the most 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.

The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

**Solution Summary**

***Step 1: Importing and analyzing data.***

* Import the data and analyze its data types.
* Find percentage of missing values in each column.
* Find the spread of data in the numerical columns.
* Find the count of unique categories in each column.

***Step 2: Data Cleaning***

* Replace ‘Select’ category with Nulls.
* Remove highly skewed and unnecessary columns.
* Remove columns with >= 30 % null values.
* Impute the nulls in those columns where null percentage is around 1%.
* Clean the categorical columns by removing duplicates and grouping low frequency categories into one.
* Handle the outliers by capping them.

***Step 3: Data Preparation***

* Create dummies for the categorical columns.
* Divide into train and test sets.
* Scale the numerical columns in train set using MinMaxScaler.

***Step 4: EDA***

* Performed EDA on categorical variables to find their category-wise counts wrt Conversion type.
* Plotted the boxplots of numerical attributes wrt Converted type.
* We noticed that customers who converted mostly spent more time on website.

Higher category-wise conversion rates were for:

* Lead origin as Lead Add form
* Lead source as Reference
* Last activity as SMS sent
* Current occupation as Working professional

***Step 5: Model Building***

* Initialize Logistic Regression model.
* Perform coarse feature selection using Recursive Feature Elimination and select top 15 features.
* Perform fine feature selection by removing the features with high p-values and Variance Inflation Factor (VIF).
* In the end, we had 11 features in the final model and all of them had a p-value < 0.05 and VIF < 5.

***Step 6: Model Performance Evaluation***

* Make predictions using the final model on the train set.
* Plot the ROC curve for the model. We obtained an impressive ROC curve that aligned to the top left edge of the plot and AUC = 0.87.
* Using the confusion matrix, plot the Accuracy, Sensitivity, Specificity for the model at different probability thresholds.
* Since in our case we want to identify most of the potential leads to minimise unnecessary calls to others, we would want our Recall to be HIGH.
* We know that Sensitivity = Recall so we set 0.35 as our Cut-off probability so we get a Higher Recall and fairly high Accuracy and Specificity.
* We obtained a target lead conversion rate of around 79% with our model.

***Step 7: Evaluation of model on the test set***

* Transform the numerical columns in test set using the MinMaxScaler object.
* Choose those columns from test set that are used in the final prediction model in the train set.
* Find the predictions on the test set and evaluate the performance metrics.

A picture containing text, font, screenshot, white

Description automatically generated