Lead Scoring Case Study Summary:

Problem Description:

An education company named X Education sells online courses to industry professionals. Although X Education gets a lot of leads, its lead conversion rate is very poor and is around 30%.

X Education needs help with building a logistic regression model so as 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach:

- **Reading & understanding the data:**

  - ✓ In this step we took a first look at the dataset and inspected the following:
  - ✓ First few and last few rows
  - ✓ Checked the shape of the data
  - ✓ Data types for each column
  - ✓ Got the descriptive statistics for the numerical columns
  - ✓ Did basic research to get better understanding of the domain

- **Data Cleaning:**

  - ✓ Converted 'Select' values to null values.
  - ✓ Missing value treatment:

Missing Value Treatment:

Feature Name	% of Nulls	How it was handled
How did you hear about X Education	78.5%	Dropped as missing values > 45%
Lead Profile	74.2%	Dropped as missing values > 45%
Lead Quality	51.6%	Dropped as missing values > 45%
Asymmetrique Profile Score	45.6%	Dropped as missing values > 45%
Asymmetrique Activity Score	45.6%	Dropped as missing values > 45%
Asymmetrique Activity Index	45.6%	Dropped as missing values > 45%
Asymmetrique Profile Index	45.6%	Dropped as missing values > 45%
City	39.7%	Dropped as as missing values ~40% and data skewed towards category Mumbai
Specialization	36.6%	Missing values imputed as Unknown
Tags	36.3%	Dropped as feature generated by Sales Team.
What matters most to you in choosing a course	29.3%	Dropped due to skewness towards a single category
What is your current occupation	29.1%	Replaced null values with Unknown to avoid skewing the data further
Country	26.6%	Dropped as as missing values ~40% and data skewed towards category India
Page Views Per Visit	1.5%	Imputed missing values to median, given presence of outliers and capped outliers with value at 99th percentile
TotalVisits	1.5%	Imputed missing values to median, given presence of outliers and capped outliers with value at 99th percentile
Last Activity	1.1%	Used mode email opened to impute the data and clubbed categories with lower frequency into 'Others'
Lead Source	0.4%	Converted category google to Google and also clubbed categories with lower frequency into 'Others'

  - ✓ Further dropped columns with only one unique value:
  - ✓ Dropped columns with unique values = 2, after confirming data imbalance of > 85%
  - ✓ Checked for duplicates, none were found.

- **Exploratory Data Analysis:**

  - ✓ Did basic EDA and identified very interesting patterns in the data.
  - ✓ Performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
  - ✓ Dropped the column 'Last Notable Activity' as the feature is sales team generated
  - ✓ Performed bivariate analysis on numerical columns by plotting box plots.
  - ✓ Also used a heat plot to identify highly correlated numerical columns.

- **Data Preparation:**

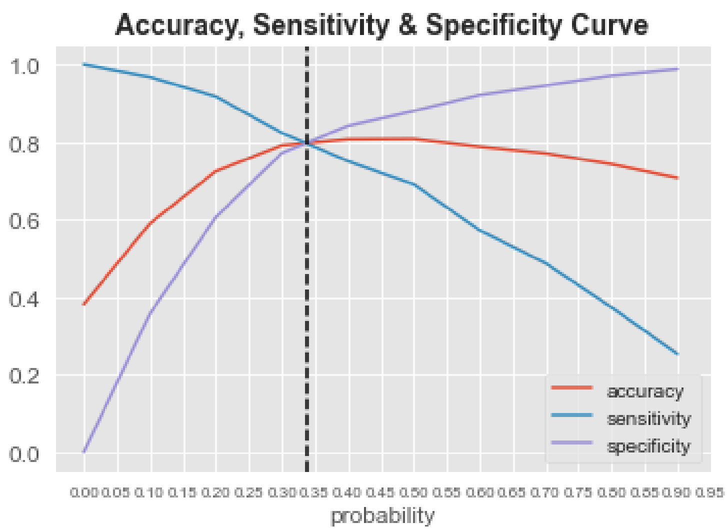
  - ✓ Created dummy variables the categorical columns with more than 2 categories using the pd.get_dummies function
  - ✓ Performed a 70-30 spilt the leads dataset into Train and Test respectively
  - ✓ Performed feature scaling using the standard scaler.

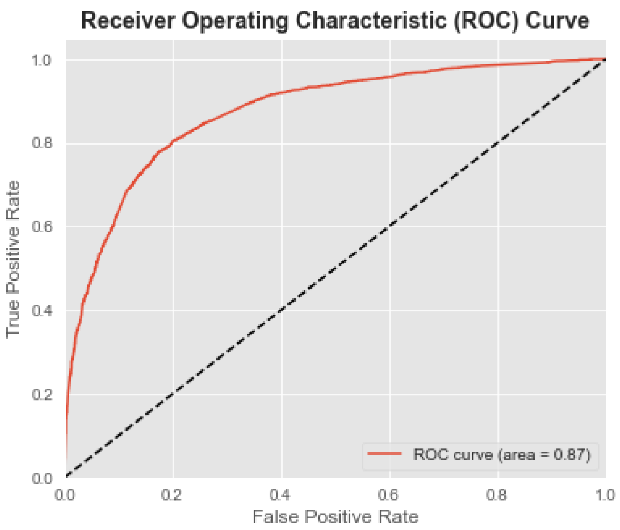
- **Model Building:**

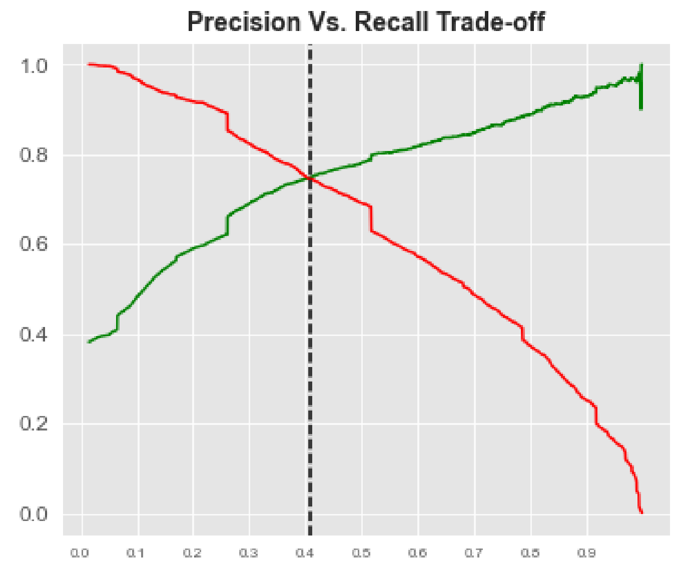
  - ✓ We shortlisted the top 15 features using the Recursive Feature Elimination (RFE) technique to build our first model.
  - ✓ In the next few iterations, we further fine-tuned our model by eliminating features with p-values > 0.05 and (Variable Inflation Factor) vif values > 5. Using vif helps reduce the impact of multicollinearity in the data.
  - ✓ Once this model was less complex with ~10 features, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0.

- **Model Evaluation:**

  - ✓ We also calculated the metrics sensitivity, specificity, precision, and accuracy.
  - ✓ To make predictions on the train dataset, optimum cut-off of 0.34 was found from the intersection of sensitivity, specificity and accuracy as shown in below figure.
  - ✓ We also plotted roc curve to find the area under the curve (0.87 for the train data set).
  - ✓ We also tired getting the optimal cut-off using Precision vs. Recall Trade-off curve. However, the models sensitivity and precision went below the 75% mark and hence was not considered in as the final cut-off.



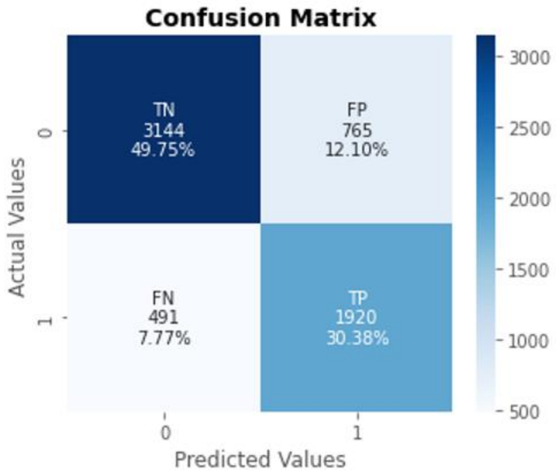





```
Model Accuracy        :  80.8 %
Model Sensitivity     :  74.5 %
Model Specificity     :  84.7 %
Model Precision       :  75.0 %
```

- **Predictions on the Test Set:**
  - ✓ After finalizing the optimum cut-off of 0.34 and calculating the metrics on train set, we predicted the data on test data set. Below are the observations:

TRAIN DATA:



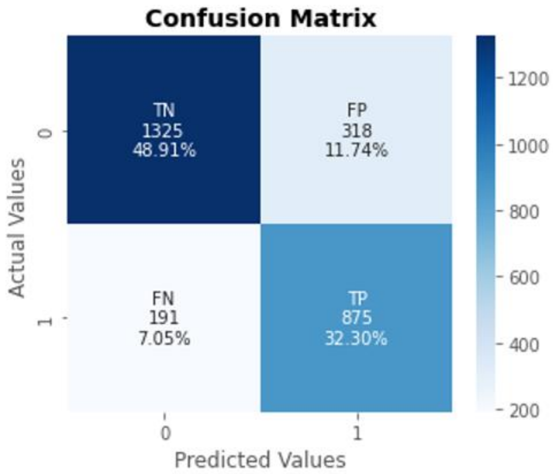

```
Model Accuracy        :  80.1 %
Model Sensitivity     :  79.6 %
Model Specificity     :  80.4 %
Model Precision       :  71.5 %

              precision    recall  f1-score   support

           0       0.86      0.80      0.83      3909
           1       0.72      0.80      0.75      2411

    accuracy                           0.80      6320
   macro avg       0.79      0.80      0.79      6320
weighted avg       0.81      0.80      0.80      6320
```

TEST DATA:




```
Model Accuracy        :  81.2 %
Model Sensitivity     :  82.1 %
Model Specificity     :  80.6 %
Model Precision       :  73.3 %

              precision    recall  f1-score   support

           0       0.87      0.81      0.84      1643
           1       0.73      0.82      0.77      1066

    accuracy                           0.81      2709
   macro avg       0.80      0.81      0.81      2709
weighted avg       0.82      0.81      0.81      2709
```

- **Final Observations:**

Below are the predictor variables that we used in our final model and their relative importance: