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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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. A typical lead conversion process can be represented using the following funnel:

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

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 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:

Step1: Reading and Understanding Data: Read and inspected the data.

Step2: Data Cleaning:

- First step to clean the dataset we chose was to drop the variables having unique values.
- Then, there were few columns with value 'Select' which means the leads did not choose any given option. We changed those values to Null values.
- We dropped the columns having NULL values greater than 50%.
- Changed NA values to Others for country names
-

Step3: Exploratory Data Analysis

- Univariant Analysis and Bivariant Analysis
- Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step 4: Building Predictive Model

- Splitting into Train and Test
- PCA Analysis
- Recursive Feature Elimination and Cross Validation

Step 5: Assessing the Model

- Fitting the Model
- Model Performance
- Finding Optimal Cutt off
- Measuring Performance on test set

Step 6: Conclusion:

• The lead score calculated in the test set of data shows the conversion rate of 83% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.

• Good value of sensitivity of our model will help to select the most promising leads.

]:

	lead_score	predicted_outcome	actual_outcome
0	6	0	0
1	58	1	1
2	86	1	1
3	12	0	0
4	19	0	0