Lead Scoring Case Study Summary:

Problem Statement:

X Education sells online courses to industry professionals. It gets leads from various sources. 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. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. The company requires you to build a model wherein you need to assign a lead score to each of the leads.

Summary:

Analysis is done on the dataset and model was built following below steps:

Reading and understanding the data.

Data loaded into pandas' data frame and observed its shape and various columns, their type and size.

Data Cleaning:

- Columns having a higher percentage of null values are dropped.
- Columns having very less percentage of null values, their corresponding rows are dropped.
- Some columns with null values are handled by imputing them with mode values.
- Few columns have 'Select' as value, which is as good as null, so handling it as null value.
- Column 'Specialization' has many categories with less values which are grouped into one category as 'Others'
- Column 'Lead Source' has Google in small and upper case which is replaced to be 'Google'.

Handling Outliers of Numerical variables using Univariate Analysis:

- Outliers of the numerical variables are visualized using univariate analysis, and with the help of quantiles, outliers are removed.

Bivariate Analysis:

- Plotted all the categorical variables using sns plots to understand the count of target variable in each category.
- Have made inferences about the lead conversion and lead origin based on the target variable count in category values of each categorical column.
- Dropped few columns from which no inference can be drawn, or the data is biased.

Dummy Variables Creation:

- Dummy variables are created for all the categorical variables and the first column is dropped.
- Original categorical variables have been dropped.

Test-Train Split:

Data is split into test and train data set for building model on train set and testing it on test set.

Scaling features:

- All the numerical variables of the train data set are scaled using standard scaler to feed to model to get proper coefficients.

Correlation:

Correlation of the variables are observed using heatmap and Corr function. Highly correlated variables are dropped.

Model Building:

- The model is built on the data frame with feature selection using RFE from sklearn with 15 variables.
- Statistics of the model are observed using statsmodels.
- models are designed by dropping variables basing on the p-value of the variables (P<0.05) and VIF(VIF<5).
- Assuming the cut off probability as 0.5. Target variable is predicted.
- Accuracy along with other metrics are calculated.
- ROC Curve is plotted to check the area of the curve and the curve is more aligned to upper part of the graph.

- Accuracy, sensitivity and specificity are calculated for different probabilities and optimal cutoff is found based on the plot of the characteristics.
- Target variable is predicted again with the optimal cut off.
- All the metrics are calculated.
- Precision and Recall have been calculated and curve is drawn.
- The model is evaluated on the test set and all the metrics are calculated and tested the model efficiency.

Below inferences are drawn from the model:

Company should focus on making calls for the leads :

- The Lead Source is Welingak Website/Olark Chat.
- Whose Lead Origin is Lead Add Form
- Whose Last Activity is a Phone Conversation/SMS Sent.
- Who spent more time on the website.
- Who are working professionals