Summary

The following analysis is done for X Education Company to get information on lead conversion. The dataset considered has around 9000 rows and 37 columns. The goal is to build a logistic regression model that gets the probability of the user getting converted to a hot lead and get information on top 100 leads. Following are the steps included in model building.

1. Data Cleaning

Few of the columns are removed that are

- not relevant to business case study
- biased to a single category.

One of the two unique identification columns is dropped i.e, 'Prospect ID'.
Select value across most of the columns is considered as null.
Drop column with highest null values ('How did you hear about X Education' 78%).
Replace null values of

- Lead Source with Mode
- Total Visits, Page Views Per Visit with Median
- Few categories are grouped as others or unknown.

2. Data Visualization

No. of converted leads (38.53) is significantly lower than non-converted (61).
Total time spent for converted leads is more than non-converted.
Working Professionals have a higher number of converted leads.

3. Data Preparation

Converted couple of binary categorical variables to 0/1
Dummification to convert categorical to indicator variables.
Split the original data into train and test data (70, 30).
Perform Standard scaling on continuous variables.

4. Model Building

Build a base model with 0.5 threshold and later perform Feature selection using RFE and select 20 variables initially.
Build a model using statsmodel for detailed statistics and eliminate variables that have either p-value (>0.05) or VIF (>5).
After elimination the total number of variables are 11.

5. Model Assessment

Model is assessed based on ROC curve, Sensitivity, Accuracy, Specificity.
For model with 0.5 threshold

- ROC curve area - 0.86
- Sensitivity - 0.668
- Accuracy - 0.808
- Specificity - 0.894

6. Optimal Threshold Value

To get a better model for predicting leads, Optimal Threshold is calculated by plotting a graph between accuracy, sensitivity, specificity. Based on the graph 0.29 is taken as the optimal threshold.

Similarly a precision recall tradeoff value is also considered for a cutoff value. Optimal value from precision recall trade off is 0.36 which is providing slightly less accurate values compared to calculated optimal threshold.

So considering Optimal Threshold of 0.29 as cutoff.

7. Model Prediction

Model built using above specified parameters is built and following are the predictions on test and train data.

Predictions taken on train data provided

- ROC Curve area - 0.86
- Sensitivity - 0.784
- Accuracy - 0.789
- Specificity - 0.792

Predictions taken on test data provided

- ROC Curve area - 0.87
- Accuracy - 0.789
- Sensitivity - 0.781
- Specificity - 0.795

8. Results and Summary

Sensitivity of trained and test models is around 0.784 and 0.781.
Lead Score is calculated for each Lead and stored in a table.
Based on the analysis following are top categorical variables in model building for better Sesitivity/Recall.

- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Lead Origin_Lead Add Form