Case Study - Lead Scoring :: Final Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. Traffic to company's website is directed through multiple channels like search engines, referrals etc. We are expected to study the data of lead generation to the final conversion (which is currently very low at 30%) and design a model which can improve the lead conversion process to more than 80% .

Approach:

Data Sanitization : EDA
After importing the data & performing the primary data scan, we applied the following data cleanup steps

- Removing columns with > 30% nulls
- Dropping null values Rows for columns like 'What is your current occupation', 'TotalVisits' etc.
- Dropped columns (like City, Country, Lead Profile etc) where count of NAN and Select (User missed selecting values) is high –
- For other columns replaced the Categorical variable 'Select' with the mode\median.
- Removed all the unwanted columns, where majority of the data was the same.
- Then data visualization was done through graphs & heatmap

Post data clean up steps, categorical variables was encoded with Dummy variable & all the redundant/repeated variables were removed.

Test Split & Model Building:

- Data was then split up (70:30 :: Train:Test) for model building.
- Feature rescaling is done to rationalize the data values at the same level.
- Since the number of variable were very high, RFE was leveraged to select top 15 variables.
- Through GLM on train data, the p-values & VIF was analyzed, and variables with high p-value/VIF were dropped one-by-one. This process was executed recursively, till all the selected variables had a VIF <5 and low p-values.
- Final Model was identified with 11 variables.
- This final model was then evaluated to calculate the Sensitivity (0.73), Specificity (0.83), CutOff (0.42, ROC (0.86 or 86%) etc
- Model created was then applied on the Test data. Metric calculated on the Test data
    - Accuracy: 78.55%
    - Sensitivity: 78.82%
    - Specificity: 78.31 %

Conclusion:

- Some Variables which have maximum impact on the Lead Score are TotalVisits, Total Time Spent on Website, Lead Origin_Lead Add Form.
- Conversion rate of 79% is very close the CEO expectation.