Summary

The scoring model building and analysis was done for the institution X Education, to find the varied possibilities to attract more industry professionals to pursue the courses offered. The dataset provided for the case study contained information about:

- Total visits
- Total Time spent on the website
- Last Activity on the site
- How they heard and reached the website
- Conversion of the leads, defining the Conversion rate

In order to build the scoring model the following approach was adopted:

Cleaning & Understanding the data

The provided has some inconsistencies as with datasets available in a real scenario. Variables with acceptable percentage of null values were removed to avoid them from influencing the analysis. As per usual practice, the null values in certain variables were replaced with the obvious or a substitute value.

Exploratory Data Analysis (EDA)

EDA was done to further understand and validate the data. The numeric values were clean with no outliers and many elements in the categorical variables had low or no significance.

Dummy Variables

As the data contained categorical variables with two or more levels, dummy variables were created to represent them in the regression equation.

Train-Test Split

Post completing the above mentioned steps, the data was split at 70% and 30% for train and test respectively.

Model Building

Recursive Feature Elimination or the RFE algorithm was used for feature selection. Top 15 relevant variables were obtained using RFE. Furthermore, variables were removed manually depending on the VIF and p-values, which was assigned a threshold of <5 and <0.05 respectively.

Model Evaluation

For the purpose of measuring performance of the model, a confusion matrix was created. Using the ROC Curve, an optimum cut-off value was arrived at to find the accuracy, sensitivity and specificity. Each of the values were approximately around 80%.

Prediction

Prediction on the test data was executed with an optimum cut-off at 0.42 with accuracy, sensitivity and specificity at 80%.

Precision-Recall

This method was used to revalidate and cut-off of 0.42 was arrived at. Precision was around 78% and recall around 77% on the test data frame.

Observation & Conclusion

Basis the analysis conducted, it was found that the variables that had relevant significance are:

- Total time spent on the website
- Total number of visits
- The source of lead
- The last activity
- Lead origin with Lead ad format
- Population where the current occupation is that of a working professional

Above mentioned factors contribute in the decision-making of the potential buyers and strategizing further on the basis of highlighted variables, X Education has a high probability of increasing their conversion rate, thereby increasing profit margins.