Lead_Scoring_Case_Study

Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. ***Cleaning data:***
We first clean the data given to us , as there are lot of missing values found, a variable having more than 33% missing values are dropped. . Identified the several columns that don't add any value are removed. As we checked for the outliers, there are no major outliers found. Values under the Lead_Source like google, Google are unified into Google.

2. EDA:
In EDA, Univariate, Bivariate and Multivariate analysis were performed. Total time spend on website greater than 15 minutes that lead had more chances to convert. Working Professionals as well as lead notified by the SMS were more likely to convert. From bivariate analysis it was found that the unemployed people who aspire for better career prospect are more likely to convert.

3. ***Dummy Variables:***
The dummy variables were created to incorporate categorical variables into logistics regression models. For numerical values we used the MinMaxScaler.

4. ***Train-Test split:***
Training Set: This is used to train our machine learning model. It contains a majority of data, 70% of the total dataset.
Test Set: This is used to evaluate the performance of our trained model. It contains the remaining portion of our data, 30% of the total dataset.

5. ***Model Building:***
Firstly, RFE is done in model to eliminate the least important feature or features with least relevance. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

6. ***Model Evaluation:***
A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy 81% , sensitivity 71% and specificity which came to be around 88% .

7. ***Prediction:***
Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy 83%, sensitivity 81% and specificity of 82%.

8. ***Precision – Recall trade-off:***
The threshold value to trade-off precision and recall is found to be 0.41. Now the Precision is 76% and recall is 80% on the test data frame.

9. Conclusion:

The following are the observations that contribute to increase in probability of lead conversion.

1. Total time spent on website greater than 15 minutes, there are more chances for

lead to convert.

2. Working Professionals are more likely to convert.

3. Leads notified by SMS are more likely to convert.

4. Lead Profile with potential lead.

Similarly Top categorical variables in the model which should be focused the most are:

1. Lead Source with element Welingak Website.

2. What is your current occupation.

3. Lead Origin with element Lead Add Form.