Lead Scoring Case Study Summary

Problem Statement:

The Objective is to get the lead conversion rate to be around 80%

Solution Summary:

1. Reading and Understanding Data.
    a. Data Cleaning:
        i. We dropped the variables that have Null Values more than 35% of the dataset.
        ii. We have imputed the missing values with median values for numerical variables and with mode for classification variables
        iii. The outliers were identified and removed.
    b. Data Analysis
        i. We have dropped the columns where there was no variance/ data imbalance.
2. EDA
    a. Univariate and Bivariate Analysis
        i. We performed the analysis using the target variables and other independent variable to see if some trend exist between them.
3. Creating Dummy Variables
    a. We went on with creating dummy data for the categorical variables.
4. Test Train Split
    a. The next step was to divide the data set into test and train sections with a proportion of 70-30% values.
5. Feature Rescaling
    a. We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.
6. Feature selection using RFE:
    a. Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features. Using the statistics generated, we used p-values and VIF values to finally arrived at the 15 most significant variables.
    b. We then created the data frame having the converted probability values.
    c. We derived the Confusion Metrics and based on accuracy, Sensitivity and Specificity at different cut off, we obtain the optimal cutoff value.
7. Plotting the ROC Curve
    a. We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 90% which further solidified the of the model.
8. Finding the Optimal Cutoff Point
    a. Then we plotted the probability graph for the ‘Accuracy’, ‘Sensitivity’, and ‘Specificity’ for different probability values. The intersecting point of the graphs was

considered as the optimal probability cutoff point. The cutoff point was found out to be 0.365

b. We could also observe the new values of the ‘accuracy=82%, ‘sensitivity=81%’, ’specificity=82%’.

9. Computing the Precision and Recall metrics

a. we also found out the Precision and Recall metrics values came out to be 77% and 77% respectively on the train data set.

b. Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.43

10. Making Predictions on Test Set

a. Then we implemented the learnings to the test model and found out accuracy to 79%, Sensitivity to 80% and Specificity to be 79%.

Recommendation

X-Education will have to mainly focus below important features responsible for good conversion rate are

1. Leads who are spending more time on website should be targeted more.
2. Company should focus on 'Working Professional' as they have higher lead conversion rate.
3. Company can focus more on Welingak website to get more leads.
4. Company should role out more 'Lead Add Form' as they have highest conversion rate
5. Leads are more vocal over SMS, should company should focus more on SMS conversions, this will result in high response rate