Leads Scoring Case Study

Summary report explaining the proceedings with the assignment and the learnings that were gathered.

This analysis is done for X Education to find ways to get more industry professionals to join their courses by converting leads into conversion. 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.

Below are the steps how we have proceeded with our assignment:

Data Cleaning & Treatment:

- Remove the redundant variables/features
- Drop columns with all unique values
- Change 'Select' values to null values
- Remove columns having more than 45% null values
- For remaining missing values, impute values with maximum number of occurrences for a column
- Change the labels names in one column with two identical label names and different format into one format

Data Transformation:

- Change the multi-category labels into dummy variables and binary variables into '0' and '1'.
- Check the outliers and remove top and bottom 1% of values
- Remove all the redundant and repeated columns

Data Preparation:

- Split the dataset into train and test dataset in ratio of 70%/30% and scale the dataset
- After this, plot a heatmap to check the correlations among the variables
- Drop variables having high correlations amongst them

Model Building:

- Create a logistic regression model with RFE count 15 after the model evaluation score like AUC
- Remove rest of the variables depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept)
- Check the optimal probability cut-off by finding points and checking the accuracy for our final model, sensitivity and specificity
- Choose a cut-off at one convergent point and predict final outcomes
- Check the precision and recall with accuracy, sensitivity and specificity for final model and the trade-offs
- Record prediction made in test set and predicted value

- Evaluate model on the test set like checking the accuracy, recall/sensitivity to find how reliable the model is
- The score of accuracy and sensitivity from our final test model was found to be in the acceptable range.
- We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads.

Conclusion:

Learnings gathered are below:

A) The training data gives ROC curve value of 0.97, Accuracy of 92.29%, Sensitivity of 91.70% and Specificity of 92.66%.

- The precision score is 88.47% and recall score is 91.70% for training data.
- The test data is giving Accuracy of 92.78%, Sensitivity of 91.98% and Specificity of 93.26%, while precision score is 89.15% and recall score is 91.98%.

B) In business terms, our model is having stability and accuracy with adaptive environment skills means it will adjust with the company's requirement changes made in coming future.

C) It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
4. When the last activity was:
    a. SMS
    b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.