Lead Scoring Case Study Summary

Problem Statement

XEducation company want us to build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Data

We have total 9240 entries of unique clients and we need to identify out of these which have the highest probability of getting converted.

We deleted a few entries throughout the data cleaning procedure, maintaining 9103 entries overall.

Approximately 38% of clients were converted, as seen in the donut figure, whereas 62% were not.

Problem Mapping

Now it is time to convert the business problem to a data science problem and develop a solution. Since our target variable is categorical in nature and has 'Yes' & 'No' labels. Thus it is a classification problem, therefore we will be using Logistic Regression algorithm and the relevant performance matrix, ie, confusion matrix for this business problem.

Solution Approach

Following are the main steps that need to be considered while developing the solution for this problem –

- Importing the required libraries
- Reading and understanding the data
- Data Cleaning
- Exploratory Data Analysis (EDA)
    - UnivariateAnalysis
    - Bivariate Analysis
- Outlier Analysis
- Data Preparation
- Splitting the Data into Training and Testing Sets
- Feature Scaling
- Model Building using Logistic Regression
- Feature selection using RFE
- Making Predictions on test sets
- Model Evaluation

EDA Insights

Categorical Variables-

- The conversion rate for the "Lead Add Form" category appears to be 93%, compared to 36% for the "Landing Page Submission" category.
- The customer arriving through the "Welingak website," "Reference," and "Google" has an extraordinarily high conversion rate.
- 40% of customers who choose to receive emails about the course end up enrolling..
- Around 63% of the clients in the "SMS sent" category have converted, and 36% of the consumers in the "Email opened" category.
- The group of "Working professionals" appeared to have the greatest conversion rate overall, which is pretty evident, followed by those who were "Unemployed."
- The conversion graph for "A free copy of mastering the interview" revealed no discernible trends or insights.

- Approximately 69% of clients in the "SMS sent" category and 36% of clients in the "Email opened" category have converted; nevertheless, the numbers are comparable to the "Last action" conversion plot.

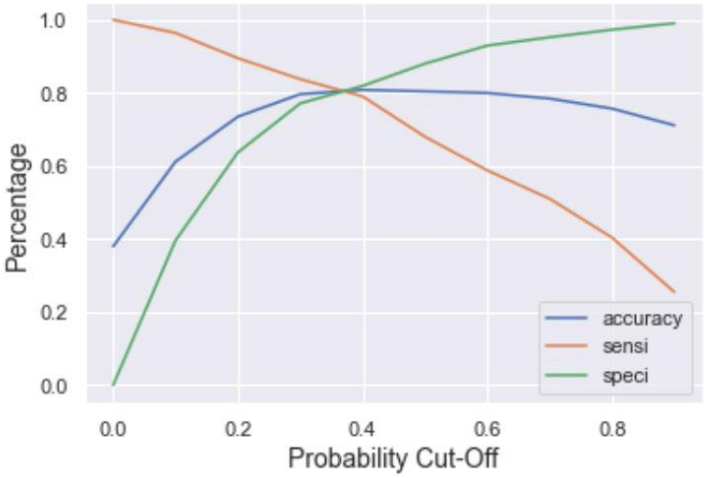
Numerical Variables-

- The average "total time spent on website" for those who converted was roughly 728, compared to 328 for those who weren't converted.
- There were no significant trends or patterns to be found in the conversion graph for "Page views per visit".

Model Building

We have used Recursive Feature Elimination technique (RFE) to remove insignificant attributes and built a model on top of that using Logistic Regression. Additionally, we ensured that the p-values and VIF values fall within reasonable parameters.

Once the stable model was created, we have predicted probabilities on the train set and evaluated performance matrics and achieved 80.5% accuracy, 80.2% sensitivity and 81.2% specificity using optimal threshold of 0.37.



Model Evaluation on Test set

Using our final model we have predicted probabilities using the same optimal threshold(0.37) on the test set and calculated accuracy, sensitivity and specificity which are as follows:

Confusion Matrix

3191	762
477	1942

- ✓ Accuracy -80.5%
- ✓ Sensitivity -80.2%
- ✓ Specificity –80.7%

Final Observation

The final model seems to predict the conversion rate very well, thus it should be able to help XEducation company in order to increase their Conversion rate to around 80%.

The top variables which are found to be very significant with respect to conversion rate are

Total time spent on Website, Working Professional, SMS Sent, Lead Add Form.

Thank you ☺