Summary – Lead scoring assignment

By Chetram Mairha, Chitwan Tayal and Chandrakant Sangam

- There are many columns which have missing values. During data cleanup SELECT in many columns are treated as missing value and all the columns above 40% missing values were removed.
- A few columns were having highly imbalanced data and those columns were dropped.
- Dropped column which are not relevant for modeling. Namely
- Dropped highly skewed columns.
- In 'Lead Source' all low Frequency variables are grouped together in 'Others' -
- In 'Last Activity' all low Frequency variables are grouped together in 'Others' –
- Mapped Binary categorical variables.
- Insights Univariate:
  - Here is the list of features from variables which are present in majority (Converted and Not Converted included)
  - Lead Origin: 'Landing Page Submission' identified 53% customers; 'API' identified 39%.
  - Current_occupation: It has 90% of the customers as Unemployed.
  - Do Not Email: 92% of the people have opted that they don't want to be emailed about the course.
  - Lead Source: 58% Lead source is from Google & Direct Traffic combined.
  - Last Activity: 68% of customers contributed in SMS Sent & Email Opened activities.
  - NOTE: These insights will be helpful in further Bivariate Analysis.
- Insights Bivariate Analysis
  - Lead Origin: Around 52% of all leads originated from 'Landing Page Submission' with a lead conversion rate (LCR) of 36%. The 'API' identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.
  - Current_occupation: Around 90% of the customers are Unemployed with lead conversion rate (LCR) of 34%. While Working Professional contributes only 7.6% of total customers with almost 92% lead conversion rate (LCR).
  - Do Not Email: 92% of the people have opted that they don't want to be emailed about the course.
  - Note: We have assumed LCR as Lead Conversion Rate in short form.

- Created Dummy variables.
- Train (70%) and Test (30%) data set created for training and testing the model.
- Did the feature scaling to bring all variables at one scale.
- Checked the co-relation and dropped highly co-related variables.
- Model Building
  - We will Build Logistic Regression Model for predicting categorical variable.
  - Feature Selection Using RFE (Coarse tuning)
  - Manually fine-tuned model using p-values and VIFs
  - In total four models built, and forth model is stable one.

- Model Evaluation done using below:
  - Confusion Matrix
  - Accuracy
  - Sensitivity and Specificity
  - Threshold determination using ROC & Finding Optimal cutoff point.
  - Precision and Recall

- Plotted ROC curve and got the intersection value 0.345.
- Did the prediction on Test dataset.

- Model result

  Train – Test

  Train Data Set:

  - **Accuracy:** 80.46%
  - **Sensitivity:** 80.05%
  - **Specificity:** 80.71%

  Test Data Set:

  - **Accuracy:** 80.34%
  - **Sensitivity:** 79.82% ≈ 80%
  - **Specificity:** 80.68%

- Recommendation to business

  Top three variables to focus on for the high probability of a lead getting converted.

  - Lead Source_Welingak Website
  - Lead Source_Reference
  - Current_occupation_Working Professional