Lead Scoring Case Study Summary

★ **Steps Followed:**

- **Data Understanding:**
    a. Checked for Null/Duplicated/Unselected Values. There are no duplicates. But found Null and unselected values.

- **Data Cleaning/Manipulation:**
    a. Removed Columns having more than 40% Null values.
    b. It was given in the problem statement that many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.
    c. 'Select' & 'NaN' with 'Unspecified' for these columns were imputed with 'Unspecified'.
    d. Null values in Numerical Columns 'TotalVisits', 'Page Views Per Visit' & Categorical Columns 'What matters most to you in choosing a course', 'What is your current occupation' were imputed with 'Mode'
    e. Dropped some unwanted columns('Country, City, Prospect ID, Lead_Number, Last Notable Activity, Do Not Call, Search, etc.) which are not useful for model building.

- **Exploratory Data Analysis:**
- **Data Imbalance Checking on 'Converted' Variable:** Successful lead conversion rate is just 38.54%. But, 61.56% of the Leads have not converted. So. the data is imbalanced.
- **Outlier Analysis:** Found that 'Total Visits' & 'Page Views Per Visit' data has outliers(high extreme values)
- **Univariate Analysis:** Observed the following from Univariate Analysis
    a. Lead Origin: The highest percentage of Leads are from 'Landing Page Submission'(52.9%) followed by 'API'(38.7%)
    b. Lead Source: The highest percentage of Leads are from 'Google'(31.5%) followed by 'API'(27.5%)
    c. Last Activity: Major Last Activities recorded are 'Email Opened'(38.3%) and 'SMS Sent'(29.7%)
    d. What is your current occupation: As per the above countplot, 89.7% of the leads are Unemployed
- **Bivariate Analysis:** Observed the following from Bivariate Analysis

a. Lead Origin: The highest leads are from "Landing Page Submission" followed by "API" Origin. Also, 36.2% of total leads from 'Land Page Submission' have been converted and 31.1% of the leads from 'API' have been successfully converted. But, the conversion rate is high for the category 'Lead Add Form'. Around 92.5% of the leads from this category have been successfully converted.
b. Lead Source: The highest leads are in "Direct Traffic" and "Google" Categories. But, Conversion rate is high in 'Welingak Website' and 'Reference' categories.
c. Do not email: 92% of the leads opted for not getting emails. Though they opted for no emails, 40.5% of these people were successfully converted.
d. Last Activity: The number of Hot leads is higher in 'SMS Sent' and in 'Email Opened' category. Conversion rate is more in 'SMS Sent' category.
e. Specialization: Most of the leads didn't specify their Specialization. Most of the remaining are from Marketing Management, HR Management & Finance Management.
f. Occupation: Though most of the leads are under the 'Unemployed' category, Businessmen and Working Professionals can be easily converted. Out of all the 3, the Conversion rate of Working Professionals is the highest.

- **Data Transformation:**
  a. Observed ‘Free Copy’ is a redundant column. So, removed.
  b. Outlier Treatment was done
  c. Changed Multi-category variables into dummies and binary variables(‘Yes’/’No’) to (‘1’/’0’)

- **Data Preparation:**
  a. Splitted the data into Train and Test Datasets.
  b. Performed Feature Scaling

- **Model Building:**

a. Used RFE for Feature Selection
b. Performed Manual Feature Elimination by removing variables whose P value is greater than 0.5 / VIF value greater than 5

- **Model Validation:**
  a. Performed Predictions on the Train & Test Datasets for 'Accuracy', 'Sensitivity', 'Specificity' whose expected range is around 80%. Obtained desirable results.
  b. .Performed probability prediction.
  c. Checked the optimal probability cut-off by points and checked the accuracy, sensitivity and specificity at this point (at 0.352).
  d. Created Confusion matrix, Accuracy, Sensitivity, and Specificity ranged in 80% (acceptable range). ROC curve (0.88 area under the curve)
  e. Performed Precision-Recall Tradeoff that gave cut off 0.422 which reduced Accuracy, Sensitivity, Specificity etc. to 75% range, So decided to use 0.352 cut-off.
  f. Assigned Lead Score on the training data.

- **Making Predictions:**

  Performed Scaling and made predictions using the final model.

- **Model Evaluation Conclusions & Recommendations:**
  a. Created Confusion matrix, Accuracy, Sensitivity, and Specificity ranged in 80% (acceptable range). ROC curve (0.88 area under the curve) on Test Model
  b. Assigned Lead Score on the training data.
  c. After running the model on the Train and Test Datasets, evaluation metrics met the expectations of X-Education's CEO, which is to achieve an 80% target lead conversion rate to be around 80%.

  Evaluation Metrics are

  Train Data:

  - Model Accuracy value : 80.47%
  - Model Sensitivity value : 79.97% ≈ 80%

- Model Specificity value : 80.78%

Test Data:

- Model Accuracy value : 80.66 %
- Model Sensitivity value : 79.73 % ≈ 80 %
- Model Specificity value : 81.28 %

Hence, our model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in reaching his goal to achieve 80% lead conversions with this model.

Recommendations:

- We can put Lead Add Forms on all social media platforms especially on 'Welingak' Website as it has positive conversion coefficient.
- We must focus on features with positive coefficients.
- Conversion rate of the leads from the Lead Source-Reference is also high. If the company offers good incentives for providing references and discounts to the converted leads using reference codes can increase the conversion rates.
- Working professionals have higher chances to convert as they can have financial stability, So more focus should be given in engaging with the Working professionals.

Thank You

By V V N V D Devi Mullapudi