Lead Scoring Case Study

V V N V D DEVI MULLAPUDI

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objectives

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Approach

Data Cleaning, Manipulation & EDA

- **Removed Columns having more than 40% Null values.**
- **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.**
- **'Select' & 'NaN' with 'Unspecified' for these columns were imputed with 'Unspecified'.**
- **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'**
- **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.**

Data Imbalance

- **Successful lead conversion rate is just 38.54%. But, 61.56% of the Leads have not converted. So. the data is imbalanced.**

Outlier Analysis

- **'Total Visits' & 'Page Views Per Visit' data has outliers(high extreme values)**

Exploratory Data Analysis

Univariate Analysis

Observations from Univariate Analysis:

- **Lead Origin: The highest percentage of Leads are from 'Landing Page Submission'(52.9%) followed by 'API'(38.7%)**

Observations from Univariate Analysis:

- **Lead Source: The highest percentage of Leads are from 'Google'(31.5%) followed by 'API'(27.5%)**

Observations from Univariate Analysis:

- **Last Activity: Major Last Activities recorded are 'Email Opened'(38.3%) and 'SMS Sent'(29.7%)**

Observations from Univariate Analysis:

- **What is your current occupation: As per the above countplot, 89.7% of the leads are Unemployed**

Bivariate Analysis

Observations from Bivariate Analysis:

- **Lead Origin:**

  The highest leads are from "Landing Page Submission" followed by "API" Origin. But, conversion rate is high for the category 'Lead Add Form'. Around 92.5% of the leads from this category have been successfully converted.

Observations from Bivariate Analysis:

- **Lead Source:**

  The highest leads are in "Direct Traffic" and "Google" Categories. But, Conversion rate is high in 'Welingkak Website' and 'Reference' categories.

Observations from Bivariate Analysis:

- **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.

Observations from Bivariate Analysis:

- **Occupation:**

  Though most of the leads are under 'Unemployed' category, Businessmen and Working Professionals can be easily converted. Out of all the 3, Conversion rate of Working Professionals is the highest.

Analysis of Numerical Attributes

Observations from Analysis:

- **'Total Visits'** data has outliers(high extreme values).
- **'Total Time Spent on Website'** has no outliers. People who spend more time on Websites can easily be converted.
- **'Page Views Per Visit'** also has outliers which need to be handled.

Observations from Analysis:

- Data is normally distributed
- **'Total Visits' & ''Page Views Per Visit'** have fair correlation b/w them.

Model Building

Steps involved:

- **Splitting the data into Train and Test Datasets**
- **Using RFE for Feature Selection**
- **Performing Manual Feature Elimination by removing variables whose P value is greater than 0.5 / VIF value greater than 5**
- **Performing Predictions on the Train & Test Datasets for ‘Accuracy’, ‘Sensitivity’, ‘Specificity’ whose expected range is around 80%**

Model Validation (Train Data)

ROC Curve

- **The Area of ROC Curve should have a value close to 1. We have got good value of 0.88. It indicates that our model is a good predictive model**

Optimal Cut-off Point

- Optimal Cut-off point is 0.352. It seems to be ideal.

Confusion Matrix & Evaluation Metrics

Our model with cut off value at 0.352 is providing an Sensitivity of 80% and Specificity is 80.78%. Sensitivity and Specificity in this case indicate how many leads the model identify correctly out of all potential leads which are converting. More than 80% is what the CEO has requested in this case study. Accuracy is 80.47%

```
Confusion Matrix:
[[3233  769]
 [ 494 1972]]
-----------------------------------------------------------
True Negatives                        :  3233
False Positves                        :  769
False Negatives                       :  494
True Positives                        :  1972
Model Accuracy value is               :  80.47 %
Model Sensitivity value is            :  79.97 %
Model Specificity value is            :  80.78 %
Model Precision value is              :  71.94 %
Model Recall value is                 :  79.97 %
Model True Positive Rate (TPR)        :  79.97 %
Model False Positive Rate (FPR)       :  19.22 %
Positive Predictive Value             :  0.7194454578620941
Negative Predictive Value             :  0.8674537161255702
```

Model Validation (Test Data)

Confusion Matrix & Evaluation Metrics

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

```
Confusion Matrix:
[[1363  314]
 [ 222  873]]
-------------------------------------------------------------
True Negatives                        :  1363
False Positves                        :  314
False Negatives                       :  222
True Positives                        :  873
Model Accuracy value is               :  80.66 %
Model Sensitivity value is            :  79.73 %
Model Specificity value is            :  81.28 %
Model Precision value is              :  73.55 %
Model Recall value is                 :  79.73 %
Model True Positive Rate (TPR)        :  79.73 %
Model False Positive Rate (FPR)       :  18.72 %
Positive Predictive Value             :  0.7354675652906487
Negative Predictive Value             :  0.8599369085173502
```

CONCLUSIONS

- After running the model on the Train and Test Datasets, evaluation metrics meet the expectations of X-Education's CEO, which is to achieve 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!

V V N V D DEVI MULLAPUDI