UpGrad

Lead Conversion Case Study

Logistic Regression Model to Predict hot leads

Rahul M
Kushagra Katiyar
Santhosh Thiyagarajan

Exploratory Data Analysis (EDA)

Data cleaning, prepration and treatment

- The Data set has 9240 rows and 37 columns
- It has both numeric and string variables as columns
- The columns with more than 45% of Data were dropped
- Columns with important data but had lot of missing values were replaced with most occuring variable
- Unnecerssary columns were dropped

Analysis from Categorical Variables

- After treating the missing values, 97% of the leads were from India
- The other categorical variables were analyzed similarly to understand their composition.
- The majority of leads are generated by Google, followed by direct traffic and Olark charts
- While the conversion rate is very high among employed people, it is higher among unemployed people.

- The majority of people come for better career prospects
- Newspaper articles, editorial forums, and newspaper and digital advertising generate similar leads and conversions.

Train-Test split & Scaling :

- Out train and test data were 70% and 30% respectively.

- We did min max scaling in the following variables ['Page Views Per Visit', 'TotalVisits', 'Total Time Spent on Website']

Model Building

- We used REF for feature selection
- REF was then performed to get the top 15 variables
- Then we manually removed the variables depending upon their REF Value and P Value.
- We created a confusion matrix and checked overall accuracy which is 80.91%

Model Evaluation

1) Sensitivity – Specificity
   On Training Data

o The optimum cut off value was found with the help of ROC curve. The area under ROC curve was 0.88.
o After Plotting the cutoff was 0.35 which gave us the following

Accuracy to be 80.91%
Sensitivity to be 79.94%
Specificity to be 81.50%.

Prediction on Test Data

We got
Accuracy to be 80.02%
Sensitivity to be 79.23%
Specificity to be 80.50%

Precision – Recall

When we do precision -Recall On Training Data

o With the cutoff of 0.35 we get the Precision & Recall of 79.29% & 70.22% respectively.
o So to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of 0.44 which gave

Accuracy was 81.80%
Precision was 75.71%
Recall was 76.32%

When we do precision -Recall On

Accuracy was 80.57%
Precision was 74.87%
Recall was 73.26%

So if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be 0.35

&

If we go with Precision – Recall Evaluation the optimal cut off value would be 0.44

CONCLUSION: TOP VARIABLES CONTRIBUTING TO LEAD CONVERSION

1) LEAD SOURCE

- Total Time Spent on Website
- Total Visits
- Direct traffic
- Google
- Welingak website
- Organic search
- Referral Sites

2) Lead Origin:

Lead Add Form

3) Last Activity:

- Do Not Email_Yes
- Last Activity_Email Bounced
- Olark chat conversation

The model was good in terms of prediction and we can definitely give a green light in using to improve the business.