Summary Report

Below approach followed to build the model for the Lead Scoring case study-

1. **Business Understanding**
   X Education company sells online courses to industry professionals. Company markets its Courses on several websites and search engines. When people land on websites, they fill form, based on that company classified as lead and past referrals. Lead conversion rate is 30% around.
   Problem Statement
   - X Education company getting lots of leads but conversion rate is very poor.

   Goal
   - Company wishes to identify "Hot" leads.
   - Build a model through which a lead score can be assigned to each lead such that customers with a higher lead score having higher conversion chance.
   - Sale team wants to focus more on communicating with Potential leads.

2. **Data Understanding**
   For this study, the given dataset and its shape, columns' data types, statistical information, null and non-null columns, and missing values percentages have been understood.

3. **Data Cleaning/Outlier Treatment**
   - Check for missing values percentage.
   - Remove the columns which are having more than 40% missing data.
   - Check each missing values column and plot it or check values count.
   - Replace "select" with more frequent value under respective columns or in some cases, retain it as a special category, under categorical features.
   - Plot "boxplot" for continuous variables to check outliers.

4. **Exploratory Data Analysis**
   - First performed uni-variate Analysis and got some inferences
   - Then bi-variate analysis is performed using count plot/pie chart
   - Multivariate using pair plot and heat map
   - From these, removed those columns which were not giving any inferences and not supporting in prediction

5. **Data Preparation**
   - Convert binary categorical values as 'yes' and 'no' into 1 and 0, respectively
   - Dummy variables are created for categorical columns
   - Feature scaling is done on continuous columns using standard scaler
   - Calculate the conversion rate which turned out to be around 38%

6. **Build**
   - Split the data as train and test sets
   - Feature selection using RFE, RFE support and RFE ranking to get top 20 variables
   - Apply logistic Binomial Regression model and get the metrics
   - Using manual method, looking for p|z| and VIF, drop columns and again run the model

7. **Run the model**
   - Using this model, predict the converted_lead in (0/1) and converted probabilities.
   - Assign a lead_id to every row.

- Now we have a dataset of lead_id, Converted_original, predicted_converted and converted probability
- Calculate the confusion matrix
- Based on it, calculate Accuracy, Sensitivity, Specificity, Precision and Recall
- Plot the ROC Curve
- Find the optimal cut-off point
- Precision and Recall trade-off based on the application

8. Prediction on test data

- Feature scaling on test data using standard scaling
- Predict the conversion lead and probability from train model
- Create a data frame with lead_id, original conversion, predicted conversion and conversion probability.
- Assign lead score by multiplying conversion probability with 100 with then round it where final prediction is 1.

9. Model Evaluation

- Calculate confusion matrix on test data
- Calculate Accuracy, Sensitivity, Specificity, Precision and Recall
- Precision tells us how accurate the model was in predicting the positive samples out of all the samples predicted to be positive. Recall tells us how accurately the model was able to identify the positive samples out of all positive samples that were actually present.
- Here, we should prefer high recall as the positive leads identified correctly need to be approached without wasting time on leads that are wrongly predicted as positive.

Conclusion

- The sales team can suggest the interns to approach ‘working professionals’ as they would want to continue education while working with the flexibility provided by online learning. At the same time, leads suggested by ‘references’ has the potential for higher conversion. So, offering more referral benefits than usual over a short duration can attract more aspirants in the said duration. Even leads through ‘Welingak Website’ would be hot leads as the traffic through this lead source is significantly high. Moreover, the tags assigned to the leads are aptly indicative that provide good support in the prediction of lead conversion. Lead profile also would provide relevant information to follow them up for positive response.
- As the utilization of time have to be optimized by the sales team, it is good to focus only on highly potential leads rather than spending time with no outcome. ‘Lead profile’ assigned to the lead would help in this as a first step. Then the ‘working professionals’ would be among hot leads as they have the capability to pay the fee for the course and would take independent decisions when compared to student community or unemployed group. Leads referred by alumni would be more promising as they would have already enquired the alumni about the course/program and have been inspired by them to take up the course for better career prospects.

Submitted by

Jyothi R
Mutyala Sridevi
Ekta Gupta