Lead Scoring Case Study

Soumyashree Behera

Vedhavathi Nanjappa

Shravani Peddigari

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%

APPROACH

1. **Data Reading and Understanding:**

   We imported the data and analysed the following features:
   - Number of rows and columns
   - Data types of each columns
   - Checking how the data is spread
   - Checking for any duplicate columns or dummy columns

2. **Data Cleaning**

   We checked for discrepancies in the dataset.
   - Checking for null values and imputing them with appropriate values
   - Columns having >40 % null values are dropped
   - We used mode imputation for categorical columns (Eg : City column)
   - Drop the columns which are not of any importance in our analysis and visualization
   - We have replaced the values of columns with low frequency into “Others” (Eg : Country column)

3. **Data Visualization (EDA)**

- Performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- Performed bivariate analysis on numerical columns with Converted column to see how the features influence Converted column.
- Used IQR method to treat the outliers in the data set.
- We have detected the outliers and treated them by removing the value above 95%ile and below 1%ile.

4. Feature Scaling

- We cleaned the data with no outliers and null values. Converted categorical values to numerical values (Eg: “Do Not Email”, “Do Not Call” columns)
- Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.
- Columns which have more than two levels have been converted to dummies using the pd.get_dummies function. Created ‘Dummy_vars’ dataset and merged this dataset with our main dataset.
- Before proceeding for model building, we have rescaled the numerical columns ('TotalVisits','Total Time Spent on Website' & 'Page Views Per Visit') by using the standard Scaler method.

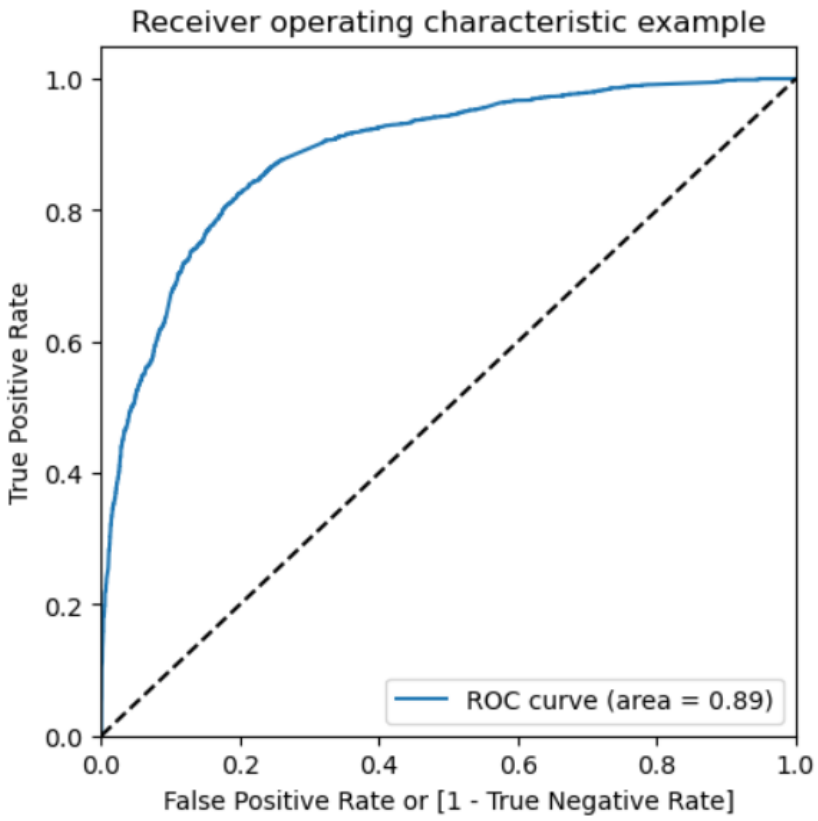
5. Model Building

- Used Recursive Feature Elimination Technique to select top 20 featured based on which model has been built. RFE uses the model accuracy to identify which attributes contribute the most to predicting the target attribute. The model stability has been evaluated by making sure features have p value < 0.05 and VIF values < 5.
- Variance inflation factor(VIF) is used to treat multicollinearity. We achieved a stable model at the end of creation of 9^{th} model.

Time:	16:28:30	**Pearson chi2:**	6.53e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.4001
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0376	0.125	-0.300	0.764	-0.283	0.208
Do Not Email	-1.5218	0.177	-8.611	0.000	-1.868	-1.175
Total Time Spent on Website	1.0954	0.040	27.225	0.000	1.017	1.174
Lead Origin_Landing Page Submission	-1.1940	0.128	-9.360	0.000	-1.444	-0.944
Lead Source_Olark Chat	1.0819	0.122	8.847	0.000	0.842	1.322
Lead Source_Reference	3.3166	0.241	13.747	0.000	2.844	3.789
Lead Source_Welingak Website	5.8115	0.728	7.981	0.000	4.384	7.239
Last Activity_Olark Chat Conversation	-0.9613	0.171	-5.610	0.000	-1.297	-0.625
Last Activity_Other_Activity	2.1751	0.463	4.699	0.000	1.268	3.082
Last Activity_SMS Sent	1.2942	0.075	17.308	0.000	1.148	1.441
Specialization_Others	-1.2025	0.125	-9.582	0.000	-1.448	-0.957
What is your current occupation_Working Professional	2.6083	0.194	13.454	0.000	2.228	2.988
Last Notable Activity_Modified	-0.9004	0.081	-11.097	0.000	-1.059	-0.741

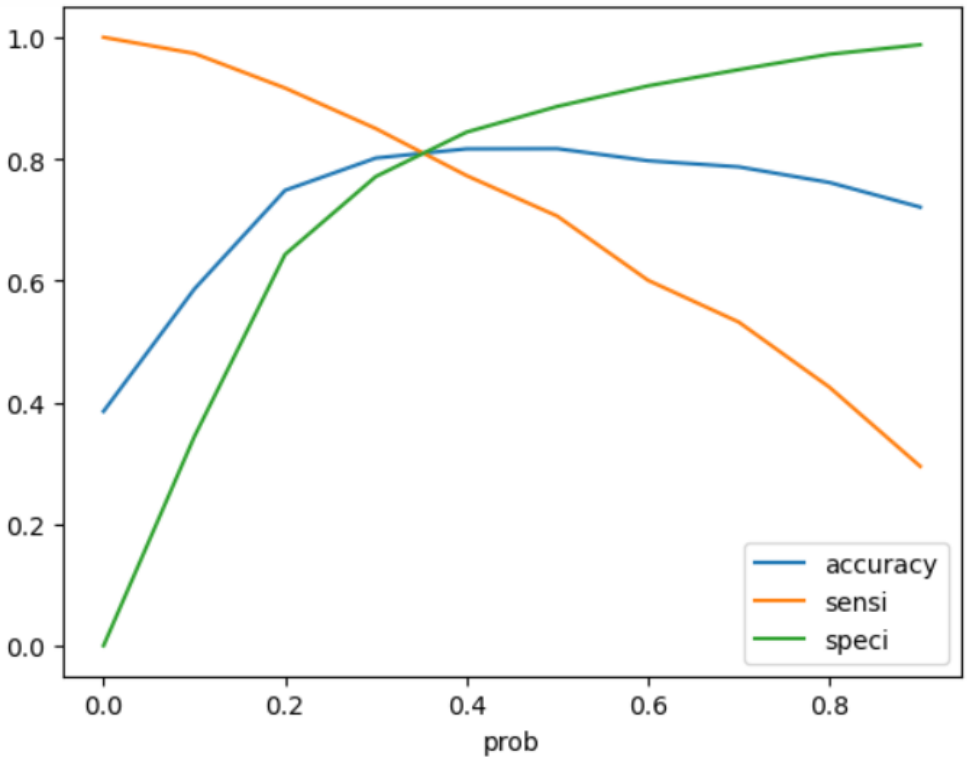
- Once the stable model was created, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0.
- We calculated the confusion matrix on this predicted column to the actual converted column. We also calculated the evaluation metrics Accuracy, Sensitivity, Specificity, Precision & Recall.
- We also plotted ROC curve to find the area under the curve (0.89 value we got).
- ROC curve shows the tradeoff between sensitivity and specificity



This model has an area under the ROC curve of 0.89, indicating that the model is good.

6. Model evaluation on Train Set

- In the previous step we took 0.5 as the cut-off. But that's not the ideal approach. Though we got Accuracy as 0.81, we need to consider Sensitivity and Specificity values.
- We calculated accuracy sensitivity and specificity for various probability cutoffs. (0.0, 0.1 …1.0)
- We plotted accuracy sensitivity and specificity for various probabilities. Based on the plot and our understanding, to make predictions on the train dataset optimal cutoff of 0.34 was chosen from the intersection of sensitivity, specificity and accuracy.



7. Predictions on Test Dataset

After finalizing the optimum cutoff and calculating the metrics on train set, we predicted the data on the test data set.
Below are the observations:

➢ **Train Data:**

- Accuracy: 81%
- Sensitivity: 81.7 %
- Specificity : 80.6 %

➢ **Test Data:**

- Accuracy: 80.4%
- Sensitivity: 80.4%
- Specificity: 80.5%

8. Final Observation:

The Model looks promising with the values obtained for the various evaluation metrics. Finally in order to convert the Hot Leads we selected Lead Score Threshold as >= 85. We identified 368 hot leads who can be contacted and have a high chance of getting converted.