Lead Score Case Study

Group Members:

Santosh Namala

Rakesh Marathe

Data source files:

Leads Data Dictionary.csv

Leads.csv

Problem Statement:

- X Education sells online courses to industry professionals.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.
- 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 Objective:

- Identify the leads that are most likely to convert into paying customers.
- Build a model wherein you need to assign a lead score 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.

Solution Methodology

- **Understanding of the data.**
- **Data Cleaning:**

  1. Check and handle duplicate data.
  2. Drop columns that: a) does not help much towards the analysis
     b) If it contains large number of missing values.
  3. Drop features that are updated by the sales team in the data as the model building is based on the data collected from the student online.
  4. Check and handle NA and missing values.
  5. Check and handle outliers in the data.

- **Splitting the data to Test and Train**
- **Model Building:** logistic regression used for building the model and RFE technique is used for feature reduction.
- **Prediction on train and test data**
- **Validation of the model- Confusion matrix , area under ROC curve**
- **Drawing conclusions and recommendations.**

Data Cleaning

- Total number of rows and columns before data cleaning are 37 and 9,740, respectively.

- Drop column lead Number as it does not help in building the model.

- As the model building is to be based on data collected from the student online, we can drop features that are updated by the Sales team :

  “Last activity”, “Last notable activity”, “Lead profile”, “Tags”, “Lead quality”, “Asymmetrique Activity Index”, “Asymmetrique Profile Index”, “Asymmetrique Activity Score” and “Asymmetrique Profile Score”.

- We observe that the percentage of missing values in few features are above the threshold level of 40 percent .Dropping the features that are greater than close to 40%.

  “How did you hear about X Education”, “Lead Profile”, “City”

- Observing the categorical data features we observe that few features are heavily skewed where the feature has more than 99 Percent data as one value, so dropping such features namely

  “Do Not Call”, “Receive More updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content” and “I agree to pay the amount through cheque”.

- Imputing missing values with median and mode wherever relevant and treating the outliers by dropping the outlier data.

Building the model

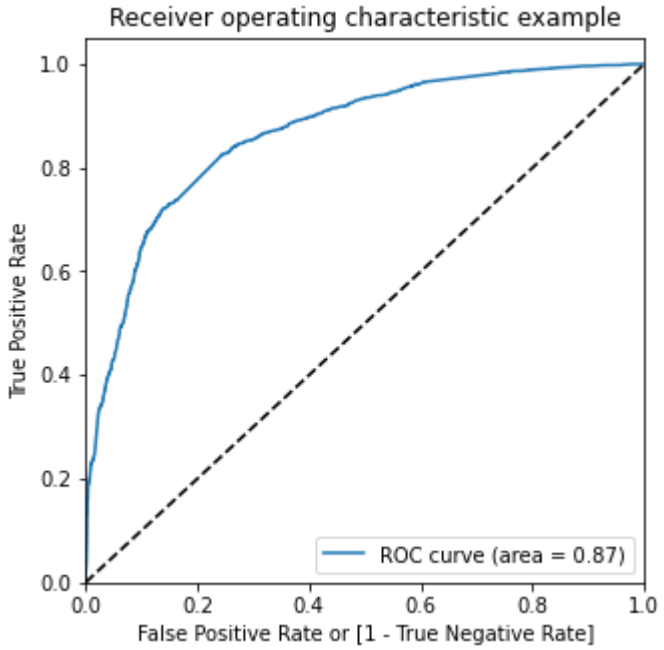
- We split the data to train and test data by 70:30.
- Model 1 is built based on the feature available after the dummy variables are created and numeric variables are scaled.
- Using RFE technique we filter the top 10 most important features .
- Based on the p-value and vif , we drop features one by one whose p-value is greater than 0.05
- After dropping features with high p-value we get our final logistic regression model with 7 important features.
-

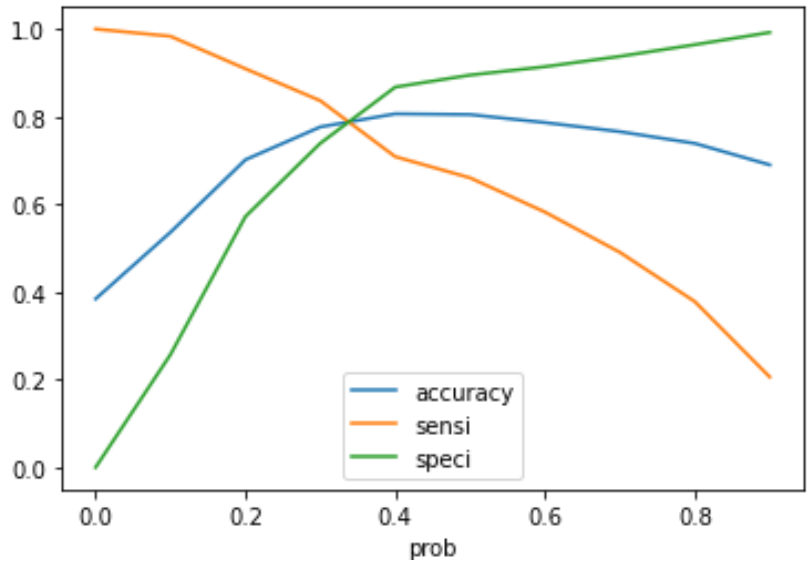
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7262	0.045	-16.278	0.000	-0.814	-0.639
Do Not Email	-1.3535	0.157	-8.596	0.000	-1.662	-1.045
Total Time Spent on Website	1.1397	0.039	29.118	0.000	1.063	1.216
Lead Origin_Lead Add Form	2.4352	0.188	12.959	0.000	2.067	2.804
Lead Source_Welingak Website	2.2047	0.739	2.983	0.003	0.756	3.654
Country_unknown	1.0224	0.095	10.732	0.000	0.836	1.209
What is your current occupation_Unknown	-1.4077	0.084	-16.680	0.000	-1.573	-1.242
What is your current occupation_Working Professional	2.3755	0.172	13.845	0.000	2.039	2.712

Model Validation

- Based on the model built, we predict the probability values on train data and randomly consider 0.3 as cut off . If the probability is greater than 0.3 it is considered as 1 (lead) and less than 0.3 it is considered as 0(not lead). Based on it we calculate confusion matrix and check the sensitivity of the model which is greater than 0.8.
- We plot the ROC curve and find the area under the ROC curve to be 0.87 which is a good value.

- ROC curve





	prob	accuracy	sensi	speci
0.0	0.0	0.384306	1.000000	0.000000
0.1	0.1	0.536449	0.983488	0.257416
0.2	0.2	0.701749	0.909384	0.572147
0.3	0.3	0.776660	0.836488	0.739316
0.4	0.4	0.806531	0.708820	0.867521
0.5	0.5	0.804829	0.660491	0.894922
0.6	0.6	0.786566	0.582360	0.914027
0.7	0.7	0.765826	0.490536	0.937657
0.8	0.8	0.739050	0.378172	0.964304
0.9	0.9	0.690141	0.206202	0.992207

Then we calculate the accuracy sensitivity and specificity for various probability cutoffs and try to plot accuracy sensitivity and specificity for various probabilities and observe the curves meet between 0.2 and 0.4. This implies our initial assumption of 0.3 cutoff is right.

Logistic Regression Model Built is :

- Target Variable = -0.7262 - 1.3535(Do Not Email) + 1.1397(Total Time Spent on Website) + 2.4352(Lead Origin_Lead Add Form) + 2.2047(Lead Source_Welingak Website) + 1.0224 (Country_unknown) -1.4077(What is your current occupation_Unknown) + 2.3755(What is your current occupation_Working Professional)

- We try to predict the test values based on the above model and we get accuracy of 0.75 and sensitivity of 0.82.

Conclusion:

The important features are:
1. Do not Email
2. Total Time Spent on Website
3. Lead Origin_LeadAdd Form
4. Lead Source_Welingak Website
5. Country Unknown
6. What is your current occupation_Unknown
7. What is your current occupation_Working Professional

In which the features Do Not Email and What is your current occupation_Unknown are negatively correlated and rest of the features are positively correlated to the target variable.

Which implies that the company must focus on leads generated from following for high lead conversion:

1. **Total time spent on Website**
2. **Lead Origin_LeadAdd Form**
3. **Lead Source_Welingak Website**
4. **What is your current occupation_Working Professional**