LEAD SCORING ASSIGNMENT

MOHD ZAID
ASHISH DIXIT
AGUM MALIK

The Summary

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

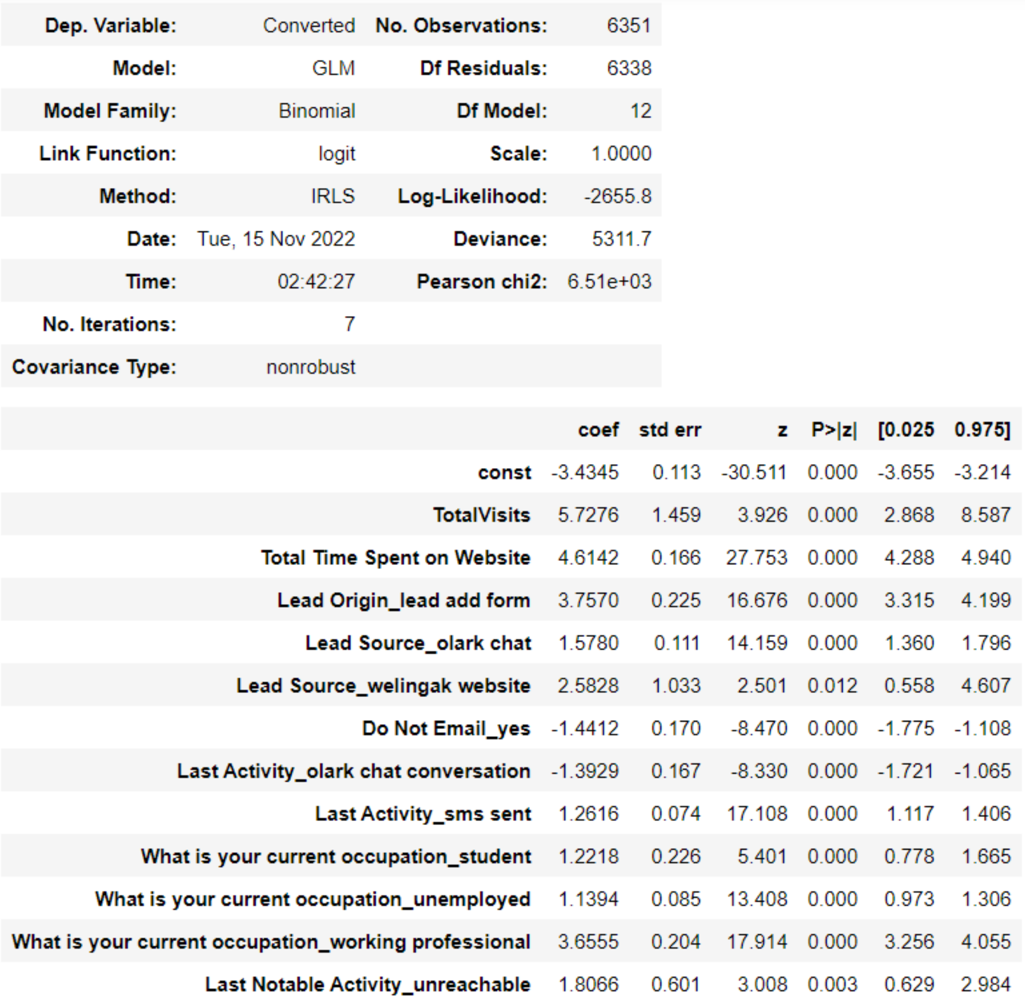
1. **EDA:**
   - Quick check was done on null values and we dropped columns with more than 45% missing values.
   - We also saw that the rows with the null value would cost us a lot of data and they were important columns. So, instead we replaced the NaN values with 'not provided'.
   - We imputed all not provided values with India as it was the most common occurrence in non-missing values.
   - We saw the Number of Values for India were quite high (nearly 97% of the Data), so this column was dropped.
   - Also worked on numerical variable, outliers and dummy variables.

2. **Train-Test split & Scaling :**

   - The split was done at 70% and 30% for Train and Test data respectively.
   - We did min-max scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

3. **Model Building**

   - RFE was used for feature selection.
   - RFE was done to attain the top 15 relevant variables.
   - Later the few of the variables were removed manually depending on the VIF values and p-value.
   - A Confusion Matrix was created, and overall accuracy recorded as 81%.



Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2655.8
Date:	Tue, 15 Nov 2022	**Deviance:**	5311.7
Time:	02:42:27	**Pearson chi2:**	6.51e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-3.4345	0.113	-30.511	0.000	-3.655	-3.214
TotalVisits	5.7276	1.459	3.926	0.000	2.868	8.587
Total Time Spent on Website	4.6142	0.166	27.753	0.000	4.288	4.940
Lead Origin_lead add form	3.7570	0.225	16.676	0.000	3.315	4.199
Lead Source_olark chat	1.5780	0.111	14.159	0.000	1.360	1.796
Lead Source_welingak website	2.5828	1.033	2.501	0.012	0.558	4.607
Do Not Email_yes	-1.4412	0.170	-8.470	0.000	-1.775	-1.108
Last Activity_olark chat conversation	-1.3929	0.167	-8.330	0.000	-1.721	-1.065
Last Activity_sms sent	1.2616	0.074	17.108	0.000	1.117	1.406
What is your current occupation_student	1.2218	0.226	5.401	0.000	0.778	1.665
What is your current occupation_unemployed	1.1394	0.085	13.408	0.000	0.973	1.306
What is your current occupation_working professional	3.6555	0.204	17.914	0.000	3.256	4.055
Last Notable Activity_unreachable	1.8066	0.601	3.008	0.003	0.629	2.984

4. Model Evaluation

- **Sensitivity – Specificity**

  We calculated Sensitivity- Specificity Evaluation:

  - On **Training Data**

    - The optimum cut off value was found using ROC curve. The area under ROC curve was 88%.
    - After plotting, we found that optimum cut-off was **0.35** which gave.

      Accuracy 80.02%
      Sensitivity 80.41%
      Specificity 80.26%.

        - Prediction on **Test Data**

            - We got:

              Accuracy 80.94%
              Sensitivity 81.51%
              Specificity 80.61%

- **Precision – Recall:**

  We calculated Precision – Recall Evaluation:

    - On **Training Data**

        - With the cut-off of 0.35 we get the Precision & Recall of 78.89% and 69.54% respectively.
        - To increase the above values, we worked on cut-off value. After plotting, we found the optimum cut off value of **0.41** which gave:

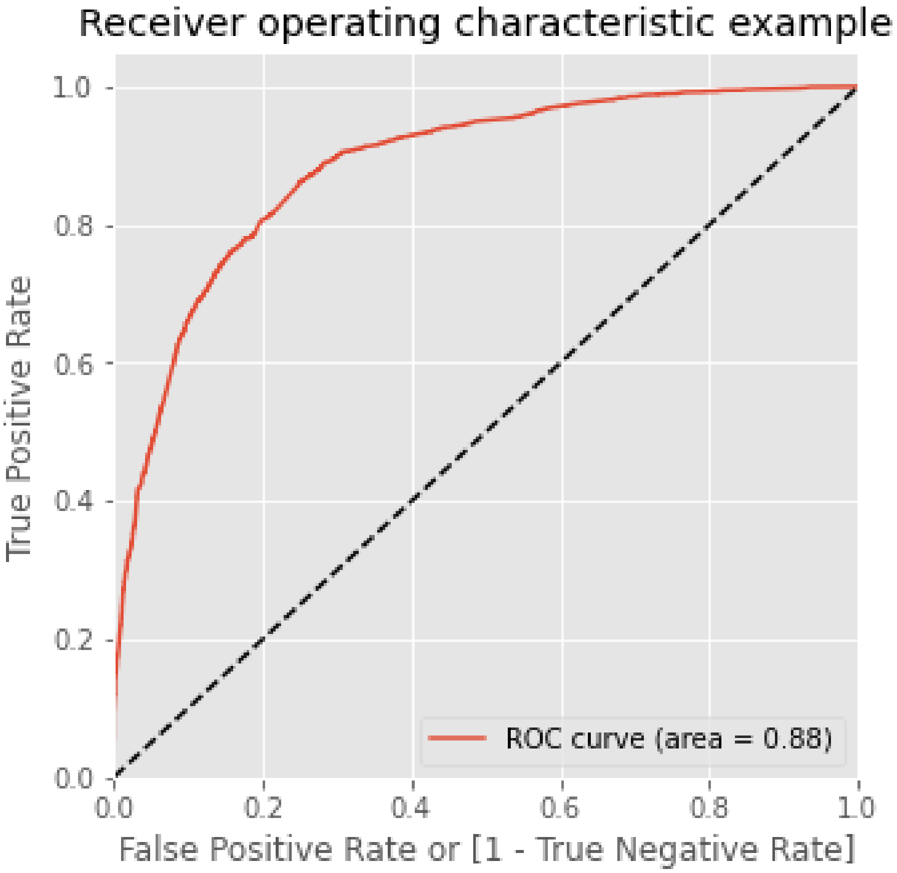
          Accuracy 81.12%
          Precision 75.45%
          Recall 75.85%

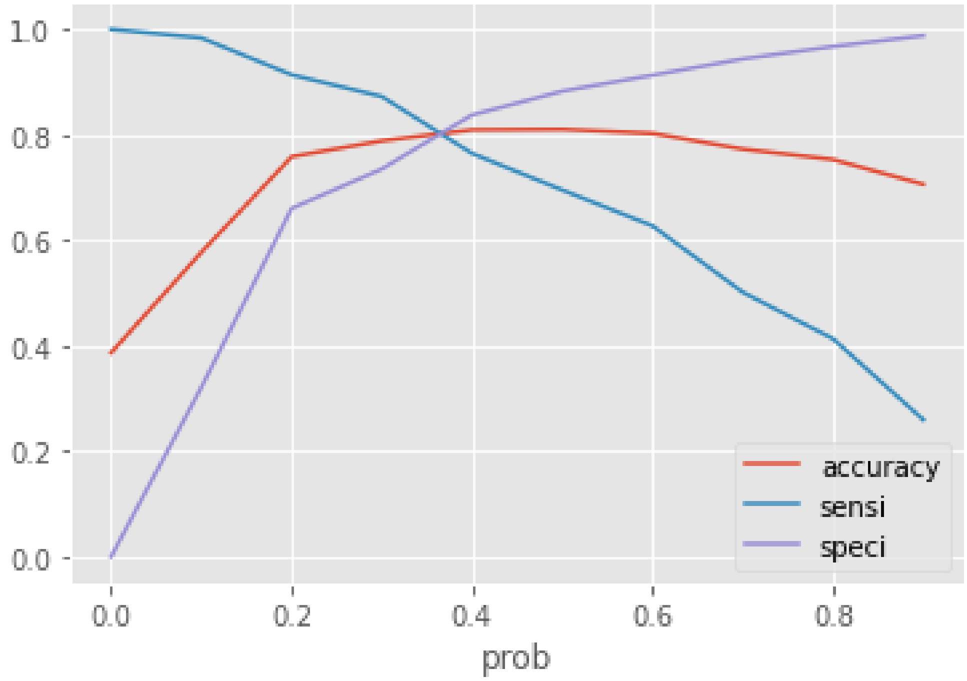
    - Prediction on **Test Data**

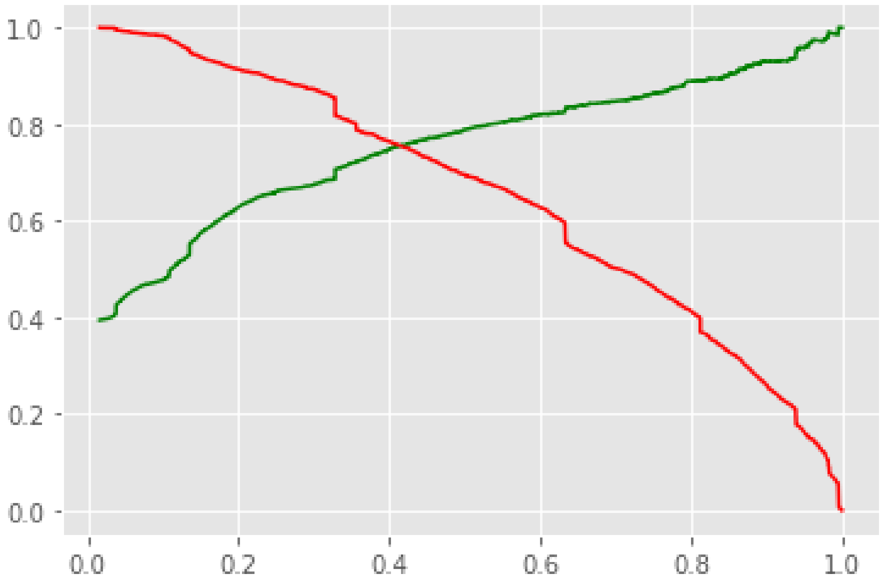
        - We got:

          Accuracy 81.49%
          Precision 73.30%
          Recall 76.30%

5. So, if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be **0.35**
And,
If we go with Precision – Recall Evaluation the optimal cut off value would be **0.44**


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.88)


0.0
0.2
0.4
0.6
0.8
1.0
prob
accuracy
sensi
speci



THE CONCLUSION

The Top Variables contributing to conversion are:

- ➢ Lead Source:
  - Total Time Spent on Website
  - Total Visits
- ➢ Lead Origin:
  - Lead Add Form
- ➢ Lead source:
  - Google
  - Direct traffic
  - Organic Search
  - Welingak website
- ➢ Referral Sites Last Activity:
  - SMS
  - Olark chat conversation

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.

The End