Assignment Summary

By- Sabyasachi Parida

A. 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%.

There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating, etc.) to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead scores have a higher conversion chance and the customers with lower lead scores have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

B. Goals of the Case Study:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

C. The following are the steps carried out:

1. **Read and inspect the data:**
   We imported the data and inspected it to understand its spread and datatypes.

2. **Data Cleaning, Outlier analysis, and Outlier Treatment:**
   - We started with imputing the **Select** Values by null as they were not selected by the lead.
   - Then went ahead with the null percentage check.
   - Deleted the columns which had more than 45% nulls.
   - Grouped categorical values, which were present in fewer numbers as **Others.**
   - In the case of Numerical variables, we imputed the null values with Median and capped the upper bound outliers to 95%.

3. **Exploratory Data Analysis (EDA):**
   We have performed a comprehensive univariate and bivariate analysis for all the variables (Categorical and Numerical) and dropped the columns which were skewed and had a high imbalance. We also dropped those Columns which were updated by the sales team.

   The following were in inferences from the Univariate analysis:

   - Most of the lead's origin is Landing page Submission (52.9%)
   - Most of the Lead source(31.1%) is from Google
   - In the case of Do not email, about 92% of people have selected No.
   - In case of Do not call 100% have selected NO.
   - Email opened is the highest in the case of the Last activity.
   - In the case of Last Notable Activity modified had 36.9% followed by 30.6%
   - 98.4% of the leads are from India.
   - In the case of specialization, Not Specified has the highest count followed by Others followed by Finance Management.
   - Most of the leads are unemployed, and about 29.1% did not select any option.
   - Most of the leads are here for better career prospects.
   - In the case of search, 99.8% selected No.
   - 100% selected No in the case of Magazine, Newspaper article, X Education Forums, NewsPaper, Digital Advertisement, 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.
   - 0.1 % leads came through Recommendations
   - Most of the Leads didn't select the city. Mumbai had the highest count after that.
   - 68.7% opted for A free copy of mastering The interview.

   The following are the inferences from the bivariate analysis :

   - Landing page submission has the highest conversion rate as well as a non-conversion rate.
   - Most of the converted leads have google as a source.
   - About 38% of leads who selected do not call or email got converted.
   - SMS sent activity had the highest conversion rate.
   - 26% of the leads which were converted were unemployed and 34 % of the leads were here for better career prospects.
   - About 39% of leads who selected NO to magazine, Newspaper, etc got converted.
   - SMS sent was the last notable activity that got converted which is about 17%.

4. **Data Preparation:** As part of data preparation, we have created the dummy variables for all categories variable after dropping the variables at EDA steps, and also performed the standard scaling to convert the numerical variables.

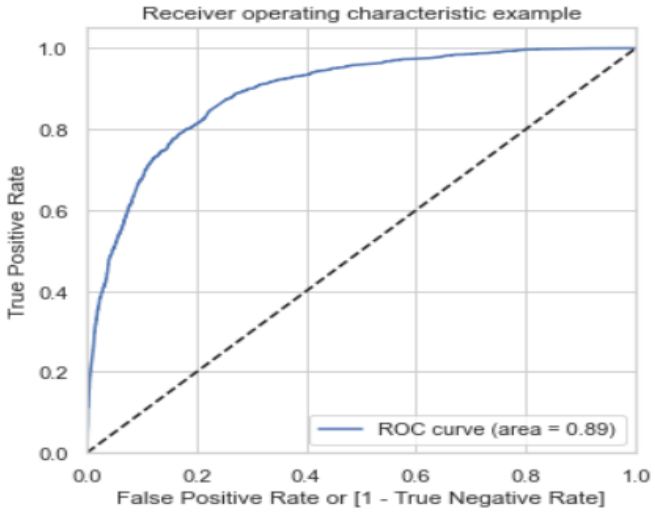
5. **Model Building:**
   We performed the mixed model approach in logistic regression. In the model building process, we **performed** the Recursive feature elimination (RFE) with initial 20 variables and then performed the manual logistic regression approach i.e. (P-value/VIF), and eliminated some variables based their p-values/VIFs and also eliminated based on their importance in the model. Finally, we arrived with 15 variables and archived a highly promising result in both train and test data sets. We observed the all variables p-values <= 0.001 and VIFs < 5.0.

6. **Model Evaluation Matrix:**

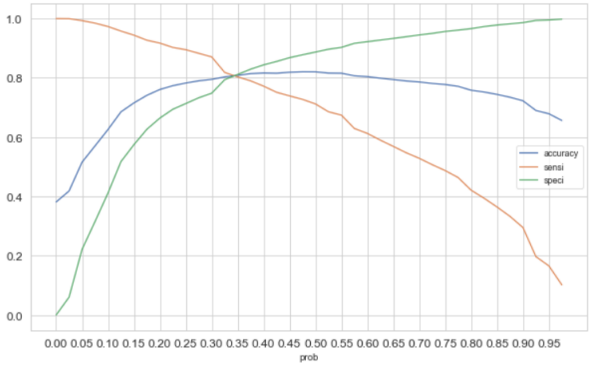
Accuracy: 82.03
Sensitivity: 71.21
Specificity: 88.71
False positive rate: 11.29
Positive predictive value: 79.53
Negative predictive value: 83.33
Precision: 79.53
Recall: 71.21

7. **ROC Curve:**

A ROC curve demonstrates several things:
- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



8. **Finding Optimal Cutoff Point:**



The optimal cutoff was found to be 0.34

9. Finding Optimal Cutoff Point

After finding the optimal cutoff the following were the metrics.
Accuracy: 82.03
Sensitivity: 80.82
Specificity: 80.53
False-positive rate: 19.47
Positive predictive value: 71.9
Negative predictive value: 87.2
Precision: 71.9
Recall: 80.82

10. Making predictions on test data

After making the predictions the following are the metrics.

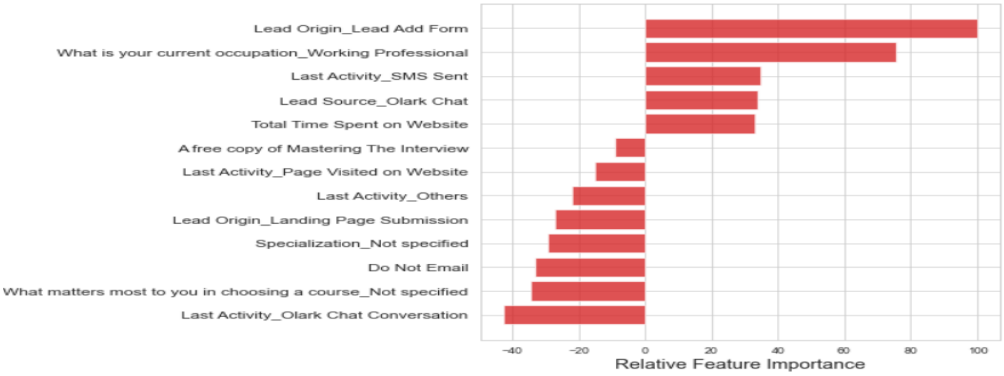
Accuracy: 80.41
Sensitivity: 80.09
Specificity: 80.62
False-positive rate: 19.38
Positive predictive value: 72.96
Negative predictive value: 86.11
Precision: 72.96
Recall: 80.09
The area under the curve: 0.8904

11. Creating the Final dataset & adding the lead score

The lead score was calculated with the following formulae

Lead Score = 100 * ConversionProbability

12. Determining Feature Importance of Final Model



Variables	Relative coeffient value
Lead Origin_Lead Add Form	100.00
What is your current occupation_Working Profes...	75.34
Last Activity_SMS Sent	34.73

With the above 3 most influencing variables.

D. Conclusion:

Overall, we produced a promising model with more 82.03% accuracy and 80.82% sensitivity in train and test datasets. As part of our systematics evaluation process steps, we started with data cleaning, then Imputed the missing values, medium capped the outliers, standardized the variables, dummified the categorical variables, and performed the mixed model approach in logistic regression. In the model building process, we performed the Recursive feature elimination (RFE) with initial 20 variables and then performed the manual logistic regression approach i.e. (P-value/VIF), and eliminated some variables based their p-values/VIFs and also eliminated based on their importance in the model. Finally, we arrived with 13 variables and archived a highly promising result in both train and test data sets. We also evaluated the other matrices and plotted the ROC curve to find out the trade-off between sensitivity and specificity. We have also performed the optimization cut-off at 0.34 and generated all potential matrices for the models. We successfully predicted the lead score and assigned it to the respective leads. We also identified the 'Hot Leads' which has a score of more than 80%. Below is a snapshot of hot leads:

	LeadID	Converted	Conv_Prob	final_predicted	Lead_Score	Lead Number
2357	3723	1	1.00	1	100	637044
1000	2764	1	1.00	1	100	649442
1880	8106	1	1.00	1	100	641324
794	5784	1	1.00	1	100	652104
3616	8080	1	1.00	1	100	625850
...	...	...	...	...	...	...
3810	8100	1	0.80	1	80	623548
3441	2923	1	0.80	1	80	627354
4207	7737	0	0.80	1	80	619506
4699	4926	1	0.80	1	80	614708
356	8197	1	0.80	1	80	656644