Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
The company needs a model wherein you 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

From above problem description we conclude that the above problem is the classification problem, hence we choose logistic Regression to calculate the Lead rate.

Below are the steps followed to solve this problem

1. **Data Reading and Understanding:**

Here we tried to get the look and feel of the data, we observed following things

- Number of rows and columns
- Data types of each columns
- Checking first few rows how data looks
- Checking how the data is spread.
- Checking for duplicates, if any.

2. **Data Cleaning:**

Here we checked for discrepancies in the dataset

- Checking for any column names correction
- Checking for null values and imputing them with appropriate methods
  - ✓ We used mode imputation for categorical columns.
  - ✓ We used mean imputation for numerical columns, if there is no skewness in data.
  - ✓ We used median imputation for numerical columns, if there is skewness in the data.

3. **Data Visualization and Outliers Treatment:**

- We performed univariate analysis on categorical column to see which columns makes more sense and removed those columns whose variance is nearly zero.
- We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.
- We performed univariate analysis on numerical columns by plotting box plots to see are there any outliers in the data or not.
- We performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.
- We have used IQR method to treat the outliers in the data set.
- In this step we also plotted the correlation matrix to identify the columns which are correlated.

4. **Feature Scaling**

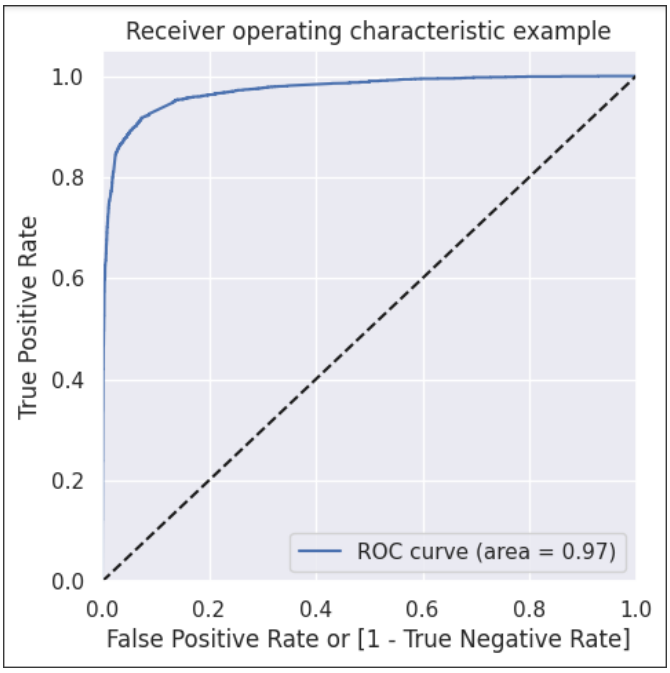
At this stage our data was very clean and no outliers. We know that logistic regression takes the input parameters as numerical values. Hence, we converted all the categorical columns to numerical.

- Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.
- Columns which have more than two levels were converted to dummies using pd.get_dummies function.

Now, the data contained only numerical columns and dummy variables. Before proceeding for model building, we have rescaled all numerical columns by using standard Scaler method.

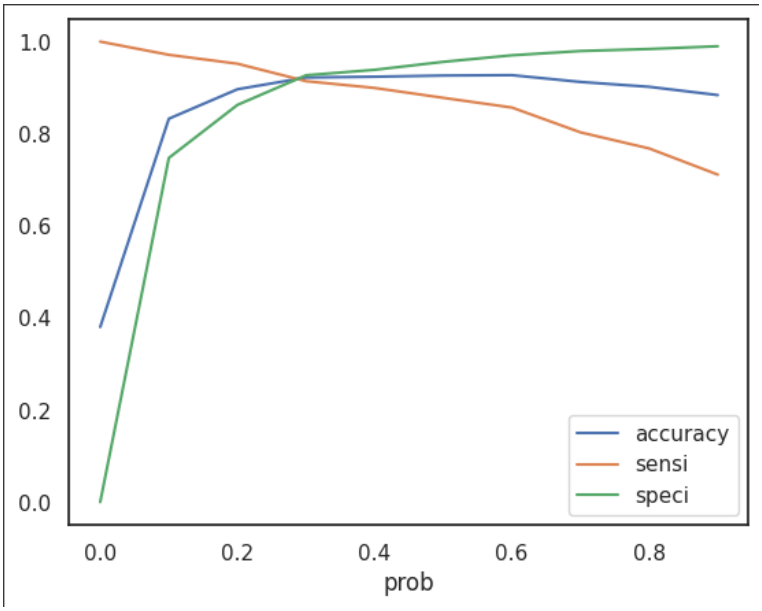
5. **Model Building**

We have used Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain. RFE uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.
In this step we made the model stable by using stats library, where we checked the p-values to be less than 0.05 and VIF values to be under 5. Variance inflation factor(VIF) is used to treat the multi-collinearity.
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 metrics sensitivity, specificity, precision, recall and accuracy. We also plotted roc curve to find the area under the curve.

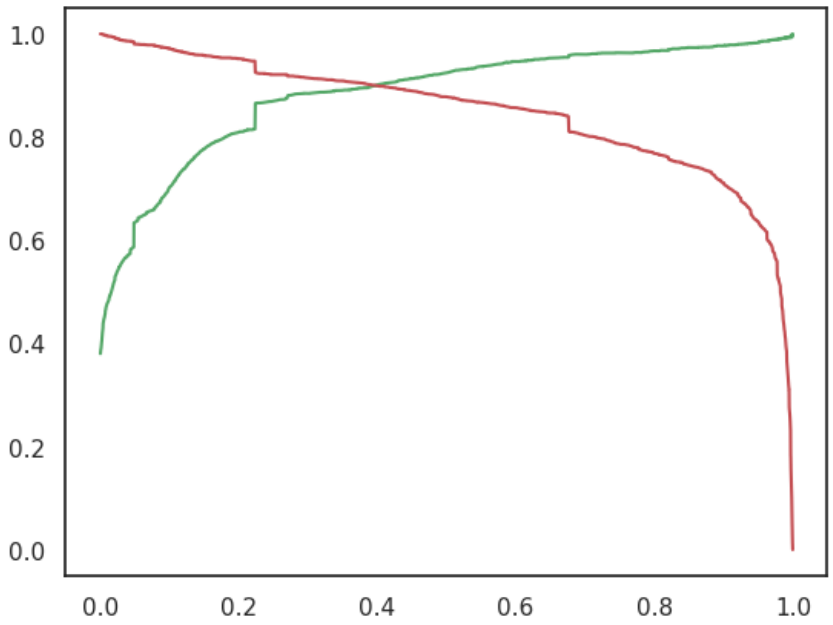


6. **Model Evaluation on Train Set**

- In the step 5 we took 0.5 as the cut-of. To confirm that it was the best cut, we calculated the probabilities with different cut-offs.
- With probabilities from 0.0 to 0.9, we calculated the 3 metrics -accuracy, sensitivity and specificity.
- To make predictions on the train dataset, optimum cutoff of 0.3 was found from the intersection of sensitivity, specificity and accuracy as shown in below figure:



- To make predictions on the test dataset, optimum cutoff was considered as obtained from Precision recall graph of the train dataset as shown below figure:



- We can observe that 0.4 is the tradeoff between Precision and Recall. Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than 40% to be a hot Lead

7. **Predictions on Test Set**

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

Train Data:

- Accuracy: 92.37%
- Sensitivity: 89.93%
- Specificity: 93.87%

Test Data:

- Accuracy: 89.09%
- Sensitivity: 95.27%
- Specificity: 85.45%

Classification Report

Train Data:

	precision	recall	f1-score	support
Not Converted	0.95	0.93	0.94	3871
Converted	0.88	0.91	0.9	2375
accuracy			**0.92**	6246
macro avg	0.92	0.92	0.92	6246
weighted avg	0.92	0.92	0.92	6246

Test Data:

	precision	recall	f1-score	support
Not Converted	0.97	0.85	0.91	1684
Converted	0.79	0.95	0.87	994
accuracy			**0.89**	2678
macro avg	0.88	0.9	0.89	2678
weighted avg	0.9	0.89	0.89	2678

8. Final Observations

The Model seems to predict the Conversion Rate very well. We should be able to help the education company select the most promising Leads or the Hot Leads.

Features which contribute more towards the probability of a lead getting converted are:

Features	**Coefficients**
Tags_Closed by Horizzon	6.59
Tags_Lost to EINS	5.49
Tags_Will revert after reading the email	4.5
Lead Source_Welingak Website	3.47
Last Activity_SMS Sent	1.98
Lead Origin_Lead Add Form	1.75
Lead Source_Olark Chat	1.23
Total Time Spent on Website	1.08
Last Activity_Email Bounced	-1.4
Last Notable Activity_Modified	-1.73
Last Notable Activity_Olark Chat Conversation	-1.85
Tags_Interested in other courses	-1.96
Tags_Other_Tags	-2.6
Tags_Ringing	-3.59

Relative feature importance are:

Features	feature_importance
Tags_Closed by Horizzon	100
Tags_Lost to EINS	83.33
Tags_Will revert after reading the email	68.25
Lead Source_Welingak Website	52.67
Last Activity_SMS Sent	30.08
Lead Origin_Lead Add Form	26.53
Lead Source_Olark Chat	18.74
Total Time Spent on Website	16.35
Last Activity_Email Bounced	-21.21
Last Notable Activity_Modified	-26.3
Last Notable Activity_Olark Chat Conversation	-28.11
Tags_Interested in other courses	-29.82
Tags_Other_Tags	-39.4
Tags_Ringing	-54.53