Summary [Lead Scoring Case Study]

Problem Description:

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, in particular, has given a ballpark of the target lead **conversion rate to be around 80%.**

Approach:

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:**

   Routine Data Check: No of rows, columns, data type of each column, distribution, mean and median for all numerical columns etc.
   Missing value analysis.
   Checking first few rows how data looks
   Checking how the data is spread.

2. **Data Cleaning:**

   "Select" value is replaced with NAN.
   Calculation of missing values for each column
   Dropping the columns with high percentage of missing values.
   Checking the unique category for each column.
   If the columns are highly skewed with one category, such columns will be dropped.
   Combining different categories of the columns with less percentage values into "Others" category.
   Imputing the column with least missing values percentage.
   Finally Checking for the number of rows kept after performing all the above steps.

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.

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.

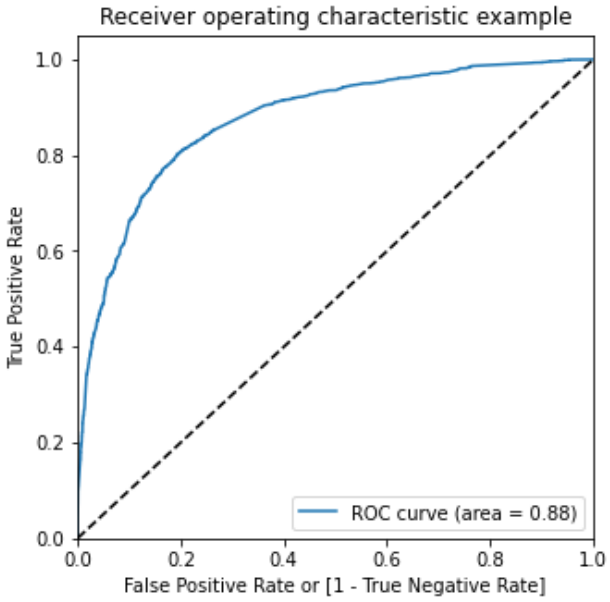
5. **Model Building:**

   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.

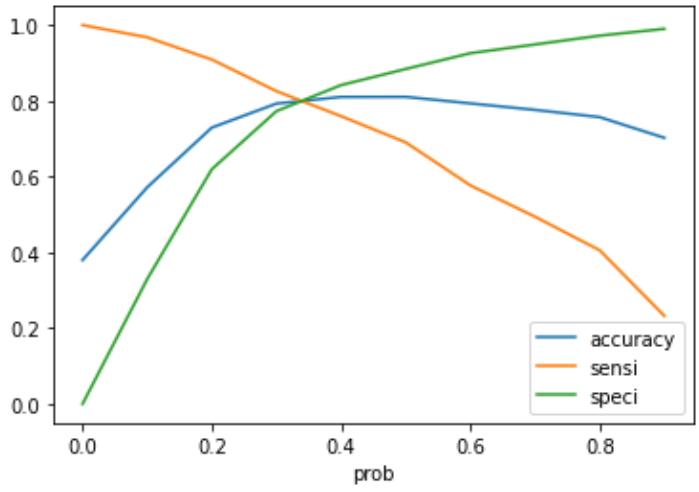


6. **Model Evaluation on Train Set**

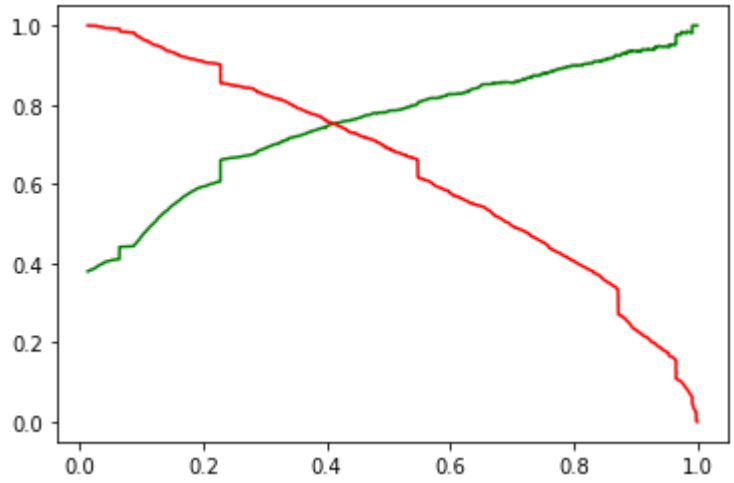
In the step 5 we took 0.5 as the cut-of. To confirm that it was the best cut, we calculated the probabilities with different cut-offs.

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 cut-off of 0.35 was found from the intersection of sensitivity, specificity and accuracy as shown in below figure:



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



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

7. **Predictions on Test Set**

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

Train Data:
Accuracy: 80.5%
Sensitivity: 79.4%
Specificity: 81.1%

Test Data:
Accuracy: 80.88%
Sensitivity: 76.1%
Specificity: 83.77%

7. **Conclusion:**

The model seems to predict the conversion rate very well and we should be able to give the CEO confidence in making good calls based on this model.
We must majorly focus on working professionals.
It's always good to focus on customers, who have spent significant time on our website.
We must majorly focus on leads whose last activity is SMS sent or Email opened.