Lead Scoring case study summary

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%.

The most promising leads, or those that are most likely to become paying customers, are chosen with the help of X Education Needs. It is necessary to create a model where each lead is given a lead score, and leads with higher lead scores have a greater chance of converting, while leads with lower lead scores have a lower chance of converting. The CEO in question has provided an approximate goal lead conversion rate of 80%.

The steps taken to solve this issue are listed below:

- **Importing the required Libraries**

1. **Here, in an attempt to represent how the data appeared and looked, we observe the following:**
    - Loading the dataset “Leads.csv” and understanding
    - Checking number of rows and columns
    - Checking the Data type of each column
    - Distribution of data
    - Check Mean and Median
    - Check missing values
    - Checking for duplicates, if any

2. **Data Cleaning:** Data cleansing is extremely important. The measures taken for data cleansing determine the model's effectiveness and quality.

    - “Select” values is replaced with NAN
    - Check number of unique values per column
    - Check for percentage of null values in each column and imputing them
    - Calculation of missing values for each column and dropping score and activity variables.
    - Dropping the column with high value of missing values
    - Finally checking for the number of rows kept after performing all the steps.

3. EDA:

In EDA, Univariate and Bi- variate analysis was done and both categorical and numerical variables.

EDA was done to check the condition of our data. It was found that many of the category factors' components were unnecessary. The numerical results seem correct, and no outliers were discovered.

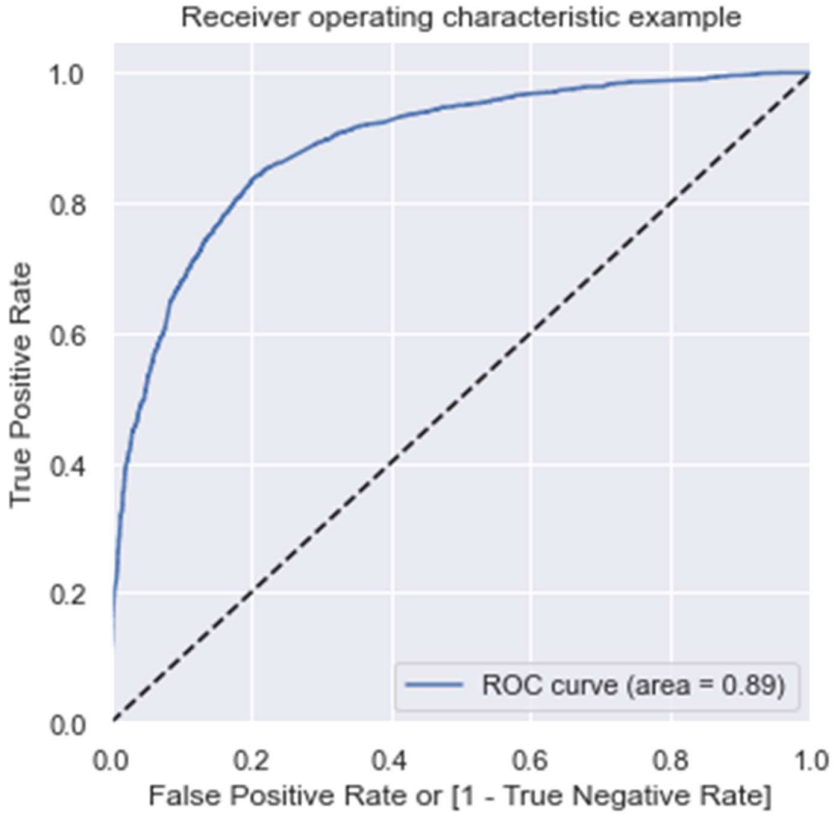
4. Model Building:

Recursive feature elimination (RFE) was used to eliminate attributes, and a model was then constructed using the attributes that remained.

To determine which characteristics are most important for predicting the target attribute, RFE looks at the model accuracy.

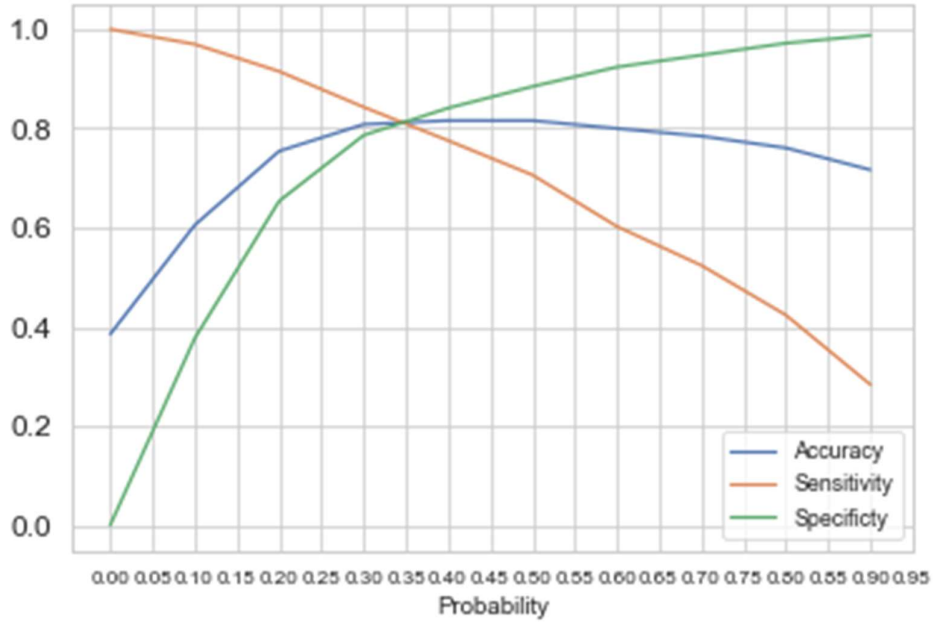
Using the stats library, we fixed the model by ensuring that the VIF values were below 5 and the p-values were less than 0.05.

We also calculated the measures for sensitivity, specificity, Precision, Recall and accuracy on this predicted column to the actual converted column. We also plotted ROC curve to find the area under the curve.



5. Model Evaluation:

- We predicted probabilities with 1 if probability is greater than 0.5 else 0.
- With probabilities from 0.0 to 0.9, we calculated the 3 metrics: Accuracy, Sensitivity and Specificity.
- The prediction on the train data set optimum cut- off 0.35 was found from the intersection of sensitivity, specificity and accuracy as shown below figure.



- We know that the relationship between in of 'y' and features variable 'X' is much more intuitive and easier to understand. The Equation is:

  log odds = 0.18 +(-1.59 Do Not Email) + (1.13 Total Time Spent on Website) + (3.49 Lead Origin_Lead Add Form) + (1.15 Lead Origin_Lead Import) + (1.08 Lead Source_Olark Chat) + (2.28 Lead Source_Welingak website) + (-0.94 Last Activity_Converted to Lead) + (-1.13 Last Activity_Email Bounced) + (1.85 Last Activity_Had a Phone Conversation) + (-1.20 Last Activity_Olark Chat Conversation) + (2.70 What is your current occupation_Working Professional) + (-1.82 Last Notable Activity_Email Link Clicked) + (-1.38 Last Notable Activity_Email Opened) + (-1.88 Last Notable Activity_Modified) + (-1.62 Last Notable Activity_Olark Chat Conversation) + (-1.67 Last Notable Activity_Page Visited on Website)

6. Conclusion and recommendation:

- Potential leads are the customers who complete out the survey.
- We must pay particular attention to leads whose most recent activity was either an email opened or an SMS sent.

- The X-Education sales staff should concentrate on prospects with lead origin- lead add form , occupation - Working Professional , Lead source - Wellingak website.
- Hot Leads are identified as 'Customers having lead score above 35. Sales Team of the company should first focus on the 'Hot Leads'
- There are many important variables like city, specialization , occupation which can potentially explain Conversion better. It is important for the management to make few of these information mandatory to fill, so that we can use in our model and build important decisions for the business.
- The 'Cold Leads'(Customer having lead score <= 35) should be focused after the Sales Team is done with the 'Hot Leads'.
- Our recall score is higher than our precision score. As a result, this strategy has the flexibility to change to meet the needs of the business in the future.
- High specificity will ensure that leads who are on the edge of being converted or not are not selected, whereas high sensitivity will ensure that almost all leads who are likely to convert are accurately forecasted.
- Customers who do not want to be contacted about the course should receive the relatively little attention.
- If the Last Notable Activity is Modified, he/she may not be the potential lead.