REPORT

This analysis is done for X Education and the company requires a model to be built to find promising leads. Lead score to be given to each lead to indicate how promising the lead could be. The higher the lead score the more promising the lead to get converted. The lower it is the lesser chances of conversion.

This could help us to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers reached the site, how frequent they visit the site, the time they spend there, their occupation, their last activity on the site, the means through which they are contacted by the employees and the conversion rate.

The following are the steps used:

1. **Data Understanding and Cleaning:**
   Firstly, we analyzed the data to know more about its Attributes.
   The data was partially clean except for a few null values and the option select had to be replaced with a null value since select basically meant that the customer did not choose anything. Few of the null values were changed to 'Not Specified' so as to not lose much data and to not create bias in data. In most categorical variables the mode was taken to replace Null values. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

   Below columns were dropped as they had null values greater than 45 percentage:
   1. Asymmetrique Activity Index
   2. Asymmetrique Profile Index
   3. Asymmetrique Activity Score
   4. Asymmetrique Profile Score
   5. Lead Quality
   6. Lead Profile
   7. How did you hear about X Education.

   After cleaning the null values we checked if any column has class imbalance and if some columns are not relevant according to business understanding. Below columns were also dropped: Country, What matters most to you in choosing a course, Search, Do Not Call, Magazine, Newspaper Article, X Education Forums , Newspaper, Digital Advertisement, Through Recommendations, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque, A free copy of Mastering The Interview.

2. **EDA:**
   A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant, Hence, we dropped it, low frequency classes in the categorical variables were combined or replaced with 'Others'. The numeric values seemed good and for some, outlier treatments were done.

3. **Data Preparation:**

   The dummy variables were created for all the categorical values and then original columns were dropped from the dataframe.

   Feature Scaling was done for numeric values as they were not of the same scale. So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. It is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. We used Standardisation (mean-0, sigma-1) for scaling.

   Train-Test split: The split was done at 70% and 30% for train and test data respectively.

4. **Model Building:**
   The model was built using the stats model and RFE was done to attain the top 15 relevant variables through feature Selection. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

5. **Model Evaluation:**
   Firstly, we calculated the accuracy of the model to check how accurately it is predecting the classes. But as we know accuracy tells you the model's performance on both classes combined which is fine but not enough. Hence it is very crucial to consider the overall business problem to decide the want to maximize or minimize. We need to create a confusion matrix to find other metrics. Using the values of this matrix we can find the following metrics:
   - Sensitivity
   - Specificity
   - Precision
   - Recall

   Now we two options, either we can maximize one metric or we can use ROC curve to find an optimal threshold at which accuracy, sensitivity and specificity are equal.

   What is the ROC curve?
   It is a trade off between True Positive Rate and False Positive Rate. A good ROC curve has a value close to 1. And our model’s value of ROC is 0.97 which is a very good value.

   After plotting the ROC curve we found that our optimal threshold is 0.3. Then we calculated the metrics at this value. They are as follows:

   - sensitivity: 91.41%
   - specificity: 92.68%

- Accuracy:92.2%

We also calculated two other metrics:

Precision: 88.46%

Recall: 91.41%

6. **Prediction on Test data:**

   Prediction was done on the test data frame and with an optimum cut off 0.3.

   - Accuracy: 92.45%
   - Sensitivity: 92.05%
   - Specificity: 92.69%

   As we can see the evaluation metrics are good when we run model on test data also. Model seems to predict conversion rate very well.

7. **Recommendations:**

   It was found that the variables that mattered the most in the potential buyers are (In descending order):

   1. When the lead source was:
      a. closed by Horizon
      b. Lost to EINS
      c. Will revert after reading the mail
   2. When the lead source was:
      d. Welingak website
      e. Direct traffic
      f. Organic search
   3. When the last activity was:
      g. SMS
      h. Olark chat conversation
   4. The total time spent on the Website.

Taking these into consideration, X Education can have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

- Target leads who belong to management specialization as this category has a high conversion rate.
- Target leads that spend a lot of time on X-education site i.e. total time spent is high.
- Target leads who are working professionals as the conversion rate is high.
- Focus on leads that come through reference.
- Avoid unemployed leads as they might not have the budget to enrol for the course.
- Avoid approaching students as they are already studying so they won't be interested but still they can be informed about the options they have for their future.