Summary

We did a machine learning analysis for X Education and find ways to get more people buying their courses. The basic data provided us with information on how they reach the website, how long the stay on the website and conversion rate.

The following steps were done to solve the ask of business 'How can we increase conversion rate'

1. Data Cleanup

The data was almost clean but had a few null values in some columns. The option 'select' in the data had to be replaced appropriately since it does not give us any additional information. We filled the null values with appropriate values as per the data column. However most of them were dropped during dummy variable creation.

2. EDA

Quick plotting of the data was done to visualize and understand our data. We found that quite a few categorical variables were not very relevant. The numeric values didnot have any outliers etc. We also dropped the columns that had more than 40% null values

3. Creation of Dummy Variables

The dummy variables were created and then we dropped the original columns. Also, We dropped dummy variables that did not provide much information based on the variance in the column.

4. Train-test split:

We split the data on standard 67:33 ratio, where Test data was 33% of whole data.

5. Building of Model:

We used RFE to gather 20 relevant features. Later we tuned the model by removing features based on VIF and p-value

6. Evaluation of Model:

We constructed confusion matrics for the model to evaluate it. We used the optimum cutoff value calculated by ROC curve and used that to find the accuracy, specificity and sensitivity of the model which was around 90% each.

7. Making Predictions

We predicted on the test data with optimum cut off at 0.41 and accuracy, sensitivity and specificity came out to be 90% each.

8. Precision and Recall

We used this method to confirm the cutoff of 0.41 and found precision and recall to be around 94% and 90% respectively

The variables that impact the potential buyers highest are below(in descending order of importance)

1. tags_will revert after reading the email
2. tags_closed by horizzon
3. last_activity_sms sent
4. total_time_spent_on_website
5. tags_lost to eins
6. lead_source_welingak website
7. last_activity_email opened

8. lead_source_direct traffic
9. lead_origin_landing page submission
10. const
11. lead_source_google
12. lead_source_organic search
13. lead_profile_student of someschool
14. tags_busy
15. lead_source_facebook