SUMMARY

X Education Company gets a lot of leads. But its lead conversion rate is very poor it is around only 30%. The company wants us to build a model to achieve a higher lead conversion score than the current rate of conversion. The CEO targets the lead score conversion rate of the company to be 80%. So, we built a logistic regression model.

- There are many columns with more than 35% of null values. These columns with more than 35% null values were dropped.
- Amongst the remaining columns ‘How did you hear about X Education’ and ‘Lead Profile’ were dropped as they didn’t have much variance.
- And these columns 'Do Not Call' , 'Search' , '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' and 'I agree to pay the amount through cheque' majorly had ‘ no ’ as their value, so they were dropped.
- The null valued rows were handled with appropriate action i.e. We dropped these rows.
- Also some of the columns that were not required were also dropped because they did not vary much. They had mostly the same value and some data columns were irrelevant.
- After cleaning, the data imbalance is checked for the target variable and it has 48% leads conversion rate.
- Performed univariate and bivariate analysis and identified the most correlated variables like ‘Last Notable Activity_Email Marked Spam’ and ‘Last Activity_Email Marked Spam’ are highly correlated to each other. Next to these were ‘Lead Source_Facebook’ and ‘Lead Origin_Lead Import’ and the third most correlated attributes were ‘Last Notable Activity_SMS Sent’ and ‘Last Activity_SMS Sent’.
- The Dummy variables were created for categorical variables
- And the Test Train split was done in 70:30 ratio.
- Feature Scaling was done using min-max scaling
- RFE was used to reduce the number of columns from 75 to 15.
- Statsmodel was used to assess the model.
- Manual Feature reduction was done from 15 columns to 11 columns. Removing about 4 columns in the final model. Feature variables with p-value greater than 0.05 was dropped.
- Initially the value of 0.5 was taken as the cutoff but the model performed poorly.
- The optimal value of the three metrics accuracy, sensitivity and specificity came at around 0.42.
- ROC was plotted and we got an area under the curve as 0.86 which indicates good predictive probability.
- The Train Accuracy was 79.08%, Sensitivity was 79.33% and Specificity was 78.84%
- For the Test Set the Accuracy was 78.45%, Sensitivity was 77.94% and Specificity was 78.91%
- As per the final model, increasing the lead conversion is very important for growth and success of X education. For this we have developed a regression model that helps us identify significant factors that impact lead conversion
- The Top 5 features with highest positive coefficient were :-
    1. TotalVisits
    2. Total Time Spent on Website
    3. Lead Origin_Lead Add Form
    4. Last Notable Activity_Unreachable
    5. Last Activity_Had a Phone Conversation