Summary

An education company named X Education sells online courses to industry professionals. They need help in selecting the most promising leads, i.e. the leads that are highly likely to convert into signing up for the courses. Requirement from the company is to build a model where a lead score is assigned to each of the leads which conveys the customer with higher and lower lead score have a higher and lower conversion chance respectively. The CEO has given a ballpark of the target lead conversion rate to be around 80%

After looking into the data, we could see that the conversion count is less, and we need to develop a model to increase this number. Which will help the company to focus on the customer who have more potential in signing up the course.

We proceeded with the EDA and found that many null values were there so we proceeded with dropping the columns with more that 40% null values. We also did some data cleaning as well. On analyzing the categorical and numerical values, we observed that:

- Quick Add form and Lead Add form seems to be the most prominent leads origin for conversion.
- Among Lead sources WeLearn, liveChat and Welingak website helped in max no of conversions.
- Denmark shows the highest rate of conversion among other countries while qatar showed the least.
- Emails showed the max conversion rate while sms were ignored.
- Amazingly max conversions are from housewives and students seemed to be a hard nut to crack.
- As per pairplot and Heat map it shows that Total votes and Page views per visit shows max correlation.
- Total time spent on website has some correlation with the person getting converted. Based on the observation we created dummy variables for the necessary categorical variables.

As we proceeded with the Model creation, we split data to test and train and later we did RFE which eliminated most of the variables from 74 and we were left with only 15 variables. After creating multiple models and dropping the high p-values attributes we selected Model 5 as the final one as the model now has 11 features, with both p-values and VIF scores within their respective thresholds.

After creating the first confusion matrix we were able to see that the sensitivity was very poor, so we changed the cut off and proceeded with the Evaluation of the model. And after multiple attempts we were able to find a cut off which gave us the evaluation metrics such that the sensitivity was high. Implying our final model didn't overfit training data and is performing well. Once metrics were calculated we assigned the Lead score to the prospect.