Lead Scoring Assignment Summary

By: Pragya Deolal & Gaurav Sawant

Problem Statement:
An education company that sells online courses to industry professionals wants to select the leads that are most likely to convert into paying customers.

Objective:
To build a model that assigns a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. Expected target lead conversion rate should be around 80%.

Approach:
The first step to begin with model building is to start by importing the required modules and dataset. The dataset initially had 9240 rows and 37 columns. Unnecessary columns and columns with missing values more than 40% are dropped from the dataset. Following the data cleaning process, data was homogenized and multicategory labels were changed into dummy variables and binary values '0' and '1'. After EDA, the dataset was split into training set and test set. LoggisticRegression() from sklearn was called to build the regression model. For model 1, Total Time Spent on Website has comparatively high p value and it also showed high VIF i.e greater than 5, thus, the variable Total Time Spent on Website was dropped for more feasible model. For model 2, the variable Total visits has comparatively high p value, thus, the variable Total visits was dropped. For model 3, the variables show decent p-value, and all the variables have VIF < 5. For the final model, the area under the curve of the ROC was 0.97 which is quite good.

Conclusion:
The accuracy estimates for the model gave promising scores in the test set as well as the training set. The burden was to work out a way to convert leads into paying customers. Initially, there were a lot of leads generated but very converted to paying customers. The model gave a high accuracy with an roc of 0.97. Important takeaways are: The variables that contributed the most towards the probability of a lead getting converted were - 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit'. Based on those hot leads could be easily figured out. It was inferred that Working Professionals are more probable to going for the course have high chances of joining it. Sources like API and Landing Page Submission brought higher number of leads and conversions. Lead Add Form had a very high conversion rate but count of leads were not as high. 'Last Notable Activity_Had a Phone Conversation' is also an important

feature. The Variable ‘What is your current occupation_Working Professional’ shows a high correlation with the probability of leads conversion whereas For Users with current occupation as Students isn't a better option to go with.