LEAD SCORING CASE STUDY

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Goals of the Case Study
There are quite a few goals for this case study.

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps of model building:

1. First, we have loaded and cleaned the data.
2. Then we have performed univariate and bivariate analysis of the features under EDA.
3. Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we dropped them.
4. Then we performed dummy variable creation for the categorical features.
5. Once data was ready, we did train test split of the data.

6. Then scaling was performed using fit_transform function of StandardScaler.
7. Automatic feature selection was done using RFE function.
8. Then we built model and done some manual feature elimination using VIF and PValue.
9. After building final model we have done model evaluation on test data where we have calculated metrics such as accuracy, sensitivity and specificity.

```
#Accuracy : 81.5 %

#Sensitivity : 68.5 %

#Specificity : 88.8 %
```

10. Below are the finally selected features of the model.

```
Lead Source_Welingak Website                                    0.6872
99
Lead Source_Reference                                           0.4493
57
What is your current occupation_Housewife                       0.3965
14
Last Activity_Had a Phone Conversation                          0.3660
47
Last Notable Activity_Unreachable                               0.2913
36
What is your current occupation_Working Professional            0.2808
66
Last Activity_SMS Sent                                          0.1968
32
Total Time Spent on Website                                     0.1833
90
Lead Source_Olark Chat                                          0.1744
25
Last Activity_Converted to Lead                                -0.0733
21
Last Notable Activity_Modified                                 -0.0932
83
Last Activity_Olark Chat Conversation                          -0.1066
45
Specialization_not provided                                    -0.1147
99
Specialization_Select                                          -0.1204
07
Lead Origin_Landing Page Submission                            -0.1349
99
Do Not Email                                                   -0.1472
88
What matters most to you in choosing a course_not provided     -0.1618
38
```

11. Recommendations of the model:

1. The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.

2. The company should make calls to the leads who are the "House Wifes" as they are more likely to get converted.

3. The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

4. The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.

5. The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.

6. The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.