Lead Scoring Assignment Summary

The case study is done for X Education and finding ways to get more industry professionals to join their cources. The basic data has been provided that gave us a lot of information about how the potential customers vist the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. Cleaning the Data:

- The data contained around 30 categorical variables
- Unique value columns removed, since they are like row number
- Geographical analysis is irrelavant in this case, since the company has to cover all over the globe the related to geographyical locations columns were dropped
- There were quite a few columns with single category which will not useful for our analysis dropped those columns as well
- Some of the columns were containing very less variety, those columns are those columns also removed
- Dropping 6 more columns which had more than 30% of null values
- The columns which had more **`“Select”`** value which is null value removed those columns as well
- Missing value treatment carried by removing missing value rows

2. Data Transformation:

- After the data cleaning started with the cleaned dataset by converting all the binary variables to ‘0’ and ‘1’ and multiple categories into dummy variables
- Next, checked the outliers of the dataset,Outliers in Logistic Regression model is very sensitive hence the outliers treated, max value into 99^{th} percentile minimum value to 1^{st} percentile

3. Data Preparation for Model building:

- Splitting the Dataset into training and test dataset with the ratio of 70:30
- Scaled the Train dataset after the splitting

4. Model Building:

- Using RFE (Recursive Feature Elimination) for automatic feature elimination
- Selected 15 feature as output from RFE
- Model has been built by removing the variable whose p-value is grater than 0.05 which is insignificant, and VIF (variance inflation factor) grater than 5 to avoid multi collinearity

5. Model Evaluation:

- Overall model predictiveness done by ROC curve got the AUC value of 0.87
- Optimal cut-off `0.425` optained by using sensitivity, specificity tradeoff
- with the optimal cut-off `0.425` we got
  - Accuracy - `0.79`
  - Sensitivity - around `0.79`
  - Specificity - around `0.79`

  - Precision - around `0.77`
  - Recall - around `0.79`
  - F1_Score - around `0.78`

6. Prediction on the Test set:

- Model prediction carried out on the test dataset with the optimal cut-off of `0.425` we got,
  - Accuracy - `0.77`
  - Sensitivity - around `0.79`
  - Specificity - around `0.79`
  - Precision - around `0.78`
  - Recall - around `0.76`
  - F1_Score - around `0.77`

7. Model Interpretation:

- The Sensitivity, Specificity, Accuracy, Precision and Recall score we got from test set in acceptable range.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.

8. Recommandations for the company:

Important features responsible for **`good conversion rate`** or the ones' which contributes more towards the probability of a lead getting converted are :

- **`Lead Source_Welingak Website`**
- **`Last Notable Activity_Unreachable`**
- **`What is your current occupation_Working Professional`**
- **`Last Activity_Had a Phone Conversation`**
- **`Lead Origin_Lead Add Form`**
- **`Last Notable Activity_SMS Sent`**
- **`Total Time Spent on Website`**

Important features responsible for **`poor conversion rate`** or the ones' which contributes lesser probability of a lead getting converted are :

- **`Last Activity_Olark Chat Conversation`**
- **`Lead Source_Google`**
- **`Lead Source_Referral Sites`**
- **`Lead Source_Organic Search`**
- **`Do Not Email`**
- **`Lead Source_Direct Traffic`**