Summary

This analysis is done on an education company named X Education which sells online courses to industry professionals, to find ways to get more industry professionals to join their courses. Our aim of this analysis to identify the “Hot Leads” so that Marketing Team can focus more on potential lead rather than communicating to all the leads.

1. Data Preparation:

- Single value columns are dropped as they don’t add value to analysis.
- Data frame contains some values as 'Select', these are replaced with null.
- Columns having more than 30% null values were dropped.
- The data is Missing Completely At Random (MCAR) as the columns are not dependent on one another
- We did check for duplicates rows.
- There are columns having many categories with very less value, a new value 'Other' is imputed for them.
- We checked for data imbalance and removed highly skewed columns.

2. EDA:

- **Univariate** and **Bivariate** analysis were done on columns to understand the data and its association with Target columns

3. Dummy Variables:

- Categorical columns were converted into dummy variables and one level was dropped

4. Normalization:

- Numerical columns were normalized using MinMaxScaler. This was done so that the coefficients of the variables are comparable and we can understand the impact and significance of one over the other to understand the most important features

5. Model Building:

- Initial top 15 features were selected using RFE and then further we build the model removing features till all features P-value was < 0.05 and VIF score was < 5

6. Model Evaluation:

- On Training data, the model evaluation generated the below stats:
    - Accuracy: 81.54%
    - Sensitivity (or Recall): 71.27%
    - Specificity: 88.08%
    - False Positive Rate: 11.91%
    - Positive Predictive Value (or Precision): 79.22%
    - Negative Predictive Value: 82.79%

- On Testing data, the model evaluation generated the below stats:
    - Accuracy: 77.96%
    - Sensitivity/Recall: 85.40%
    - Specificity: 73.50%

7. Conclusion:

Below are the variables that are most important (in order of priority) for being the potential buyers are:

- Total Time Spent on Website
- When LeadOrigin is 'Lead Add Form'
- When LastActivity was 'SMS Sent'
- When LeadSource is 'Olark Chat'
- TotalVisits
- When LastActivity was 'Email Opened'

Below are the variables that can lead a negative impact (in order of most impact) on the conversion:

- What_is_your_current_occupation has Missing Data
- What_is_your_current_occupation is 'Unemployed'
- What_is_your_current_occupation is 'Student'
- LastActivity was 'Email Bounced'
- LastActivity was 'Olark Chat Conversation'
- LeadSource was 'Direct Traffic'