Summary

The lead Score Analysis was conducted to help X Education in identification of more industry professionals to purchase their courses. This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

We took the following steps to do the complete Analysis:

✓ Data Cleaning & Manipulation

- Total Number of Rows are 9240 and columns are 37.
- Standardizing names of the columns by removing extra spaces.
- Replaced “Select” option with NaN.
- Dropped columns having for than 45% null values.
- Imputed remaining missing values of categorical variable through Mode.
- Outliers in Numerical Variable are removed by keeping to 99 Percentile Values
- Dropped columns which have very less significance as per business requirement.

✓ EDA

- We identified outliers in number of columns such as “Page_Views_Per_Visit” and “TotalVisits”
- We removed outliers by capping it to 99 percentile values

✓ Data Preparation

- Converted categorical columns "Do_Not_Email" and "A_free_copy_of_Mastering_The_Interview" with two levels(Yes and No) into numerical columns.
- Created dummy variables and dropped first and original column.
- Removed repetitive columns.
- Did Correlation Analysis on Features through HeatMap.
- Divided Dataset into train and test dataset.
- Utilized Min-max scaler for feature scaling and standardization.

✓ Model Building

- The logistic regression model is built for Prediction of lead score.
- Used RFE technique for feature selection with 15 variable as output and fined tuned it manually by checking VIF and p-Values.
- Dropped columns such as “What_is_your_current_occupation_Housewife”, “Lead_Origin_Lead Add Form”, as these features had high p values/VIF.

✓ Model Evaluation

- Evaluated the Model Performance by manually using 0.50 cut-off value.

- Using ROC Curve, trade off between Sensitivity and Specificity Parameters, established optimal cut-off value of 0.35.
- Ran Prediction for Train and Test Data Set---With the Current cut off of 0.35,

Metrics Score on Train Data Set are	**Metrics Score on Test Data Set are**
Accuracy--80.40% Sensitivity--80.38% Specificity--80.41% Precision--72% Recall-80.38% True Positive rate--75.59% False Positive Rate--19.59% Positive Prediction Value--72%% Negative Prediction Value--86.74% F1_Score--75.95%	Accuracy--80.13% Sensitivity--80.18% Specificity--80.10% Precision--69.68% Recall-80.18% True Positive rate--80.18% False Positive Rate--19.89% Positive Prediction Value--69.68%% Negative Prediction Value--87.63% F1_Score--74.56%

✓ Prediction:

- Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

✓ Summary

- It was found that the variables that mattered the most in the potential buyers are (Almost In descending order) :
  - ➢ The total time spend on the Website.
  - ➢ When the lead source was:
    a. Welingak website
    b. Reference
    c. Olark chat
  - ➢ When the last activity was:
    a. Phone Conversation
    b. SMS
    c. Olark chat conversation
  - ➢ When their current occupation is as a working professional.
  - ➢ When the lead origin is Landing page Submission

With all these Observations X Education will be easily able to identify potential leads and convert them into a successful buyer.