Lead Scoring Summary

The model building and prediction is being done for company X Education to find most promising leads. We have further understood and validate the data to reach a conclusion to target the correct group and increase conversion rate. We have followed the below mentioned steps to reach our conclusion:

1. **Cleaning data and missing value check:**
   Our first step in the process was to clean the data and perform a missing value check. We first found the unique values in the data and dropped all the columns with unique values. ‘Select’ column was replaced with “NaN” and converted all values to lower case. Then performed a missing value check and removed all the columns that are not required and have more than 35% null values. Not all the null values were removed since it will cause us a lot of data loss, so we replaced Nan values with “not provided”.

2. **EDA:**
   Then a quick EDA was performed on categorical and numeric variables to gather insights from our data. It was found that a lot of elements in the categorical variables were either irrelevant or had very less data. The numeric values seems good and no major outliers were found.

3. **Dummy Variables:**
   The dummy variables were created for ‘object’ values and later on the original column values were removed.

4. **Train-Test split:**
   The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**
   Firstly, RFE was used for selecting first 15 variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (Thevariables with VIF < 5 and p-value < 0.05 were selected). Total 12 features were selected for final model.

6. **Model Evaluation:**
   Firstly, A confusion matrix was made to check the overall accuracy which came out to be 81% which is very good value. Later on the optimum cut off value using ROC curve was used to find the sensitivity and specificity of the model which came to be around 70% and 87% respectively with a cut off of 0.5. The area under ROC curve is 0.87 which is a very good value.

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum cut off as 0.35 withaccuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**
   We used this method to recheck and a cut off of 0.41 was found with Precisionaround 73% and recall around 75% on the test data frame.

Conclusion:

It was found that the variables that mattered the most in the potential buyers are:

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:

    a. Google

    b. Direct traffic

    c. Organic search

    d. Welingak website.

4. When the last activity was: a. SMS b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

Keeping these factors in mind, the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.