SUMMARY

-By Mansi Nara

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. The following are the steps used:

- **Data Cleaning**
  The data was partially cleaned.Dropped the coulmns with >40% null values .
  There were few entries with "Select" so we replaced them with NaN values and later for some columns to avoid loss of data we replaced with "not provided".
- **EDA**
  EDA was done to check the condition of the data.Imputed some missing values in numerical columns with median and columns with unique responses were dropped.Other activities like outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out. Time spend on website shows positive impact on lead conversion.
- **Dummy Variables**
  Dummy variables were created and we dropped the columns with "not provided" . For numeric variables feature scaling was done by MinMax Scaler.
- **Train-Test-Split**
  The split was done at 70% train and 30% test respectively.
- **Model building**
  Used heat map to find correlation between the variables and the columns that and used RFE to get highly correlated columns and reduce variables to 15. This will make dataframe more manageable. Total 3 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
  The conversion rate was 39%.
- **Model Evaluation**
  Created new column 'Predicted' with 1 if Conversion_Prob > 0.5 else 0.Overall accuracy was 81% . With cutoff of 0.5 we have sensitivity around 67%,specificity around 89% and accuracy around 80%. And also got the confusion matrix.
  0.4 is the optimum point to take it as a cutoff probability.With optimum cutoff found the sensitivity, specificity,accuracy rates which were 77%,82% and 80% respectively.
- **Making predictions on Test Data**
  Converted the test data to a dataframe which is an array.
  Put Prospect ID to index.
  With cutoff of 0.41 we got the accuracy,sensitivity,specificity rate around 79%,73% and 83% respectively.
  Created confusion Matrix for the test data.
- **Conclusion**
  It was found that the variables that mattered the most in the potential buyers are (In descending order):
  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.

The most numbers of leads are from INDIA and in terms of city highest number are from Mumbai. Most of having Specialization from Finance Management. Leads from HR, Finance & marketing management specializations are high probability to convert.

Keeping these 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.