Summary

The X education sells online courses to industry professionals and they need help in selecting the most promising leads that are most likely to convert into buying customers.

The X company needs a model which can assign a lead score to each lead such that the customer with higher lead score have higher conversion rate and customer with lower lead score have lower conversion rate chance.

So, here is the strategy followed to solve this problem:

Step 1: Data Reading and inspecting

- Read the dataset provided and inspecting its size, shape and kind of variables present.

Step 2: Data Cleaning

- Handling of the 'Select' level that is present in many of the categorical variables by replacing it with the null value.
- Null values handling is done as follows,
    - Variables with null values higher than 35% are dropped from the dataset
    - Null values of categorical variable like ‘What is your current occupation’
    - replaced with mode value.
    - Removed all the rows where null values are less than 2%
- Outliers handling
    - Capping the outliers to 95% value
- Unique valued variables are also dropped from the dataset as they will not contribute for prediction

Step 3: Exploratory Data Analysis

- A brief EDA has been done by univariate analysis and bivariate analysis. Many variables are highly skewed and has unique values. Those are dropped because they are not adding any information to predict the lead conversion.

Step 4: Data preparation for model building

- Binary variables of type Yes/No are converted to 0/1
- Dummy variables are created for categorical variables and dropped repeated variables

Step 5: Train – Test Split

- The split is done into two parts called train and test with 70% and 30% of data respectively.

Step 6: Logistic Regression model building on train set

- Scaling of numeric variables in train set using min max scaler
- First model is built using 20 features selected by the RFE (Recursive Feature Elimination).
- Later rest of the features were removed depending on the p-value and VIF values. VIF values < 5 and p-value < 0.05 are kept for the model building.
- Using the final model predictions were done on the train dataset.

Step 7: Model evaluation by different measures and metrics

- Confusion Metrix is created and based on that different measures like Accuracy, Specificity and Sensitivity are calculated which is around 80%.
- Optimal cutoff value of 0.35 is found using ROC curve
- Precision and recall measures are also calculated for further verification of model efficiency.

Step 10: Testing model on test set and evaluation

- Using the final model predictions are done on test dataset
- Different measures are calculated for test dataset Accuracy (87.4%), Precision (70.4%), Recall (79.2%)

Step 11: Calculating lead score of each lead

- Lead score calculations are done for test set which shows how likely the lead can convert to potential buying customer

Conclusion:

The X company should focus on the Leads who are Spending more time on website, working professionals, Leads coming through - Reference or welingak website, whose last activity is – SMS and Email communication, whose origin is either 'Lead Add Form' or 'Lead Import, whose number of visits to website is higher.

The X company should not focus on Leads who prefers 'Do not Email’ and ‘DO not Call’, are Unemployed and Student, whose last activity was "Olark Chat Conversation".