Summary

This analysis is done for X Education company to find the most promising leads from various lead scores who can convert to potential customers. Also, a model needs to be build that can be deployed for future business usage.

Our approach :

1. Data Cleaning :
   a) No duplicate values were found in the data.
   b) The unique identifier in the data has been dropped.
   c) Missing values above 40% have been dropped.
   d) Heavily skewed variables were dropped.
   e) NaN values were imputed and merged with low frequency categories of some categorical variables.
   f) Missing values in the numerical columns were imputed by the respective medians.

2. Explanatory Data Analysis (EDA) :
   a) From the univariate and bivariate analysis of the data, some categorical columns were dropped as they added no information to the model.
   b) Outliers were found in some numerical data and top and bottom 1% of these outlier values were removed.
   c) Correlation of the numerical variables were found from the heatmap generated.

3. Dummy Variable Creation :
   a) Dummy variables were created for categorical variables with multiple levels.

4. Train-Test Split :
   a) 70% of the data was split into train set while the remaining 30% into test set.

5. Feature Scaling :
   a) The data was normalised using MinMax scaler with conversion rate of 38%.
   b) The highly correlated dummy variables were dropped.

6. Model Building :
   a) A Logistic Regression model was build using RFE feature selection to attain the top 15 relevant variables.
   b) By iterating manually, some more variables were removed based on VIF and p-values.

7. Model Evaluation :
   a) Sensitivity, specificity, accuracy were obtained from the confusion matrix.
   b) The sensitivity-specificity tradeoff was dealt using ROC curve and optimal cutoff probability.

8. Prediction :
    a) Prediction was done on the test data using optimal cutoff probability of 0.38, obtained from the precision-recall tradeoff curve.
    b) The accuracy, specificity and sensitivity of the model was more than 75%.

9. Score Variable Generation :
    a) The leads having lead score between 90 and 100 ('Hot Leads') have a very high chance of getting converted.

10. Conclusion :
    a) The most promising leads according to our model are –
        i. Working professionals
        ii. Leads who spend more time on the website and therefore the website should be upgraded with exciting offers and interesting topics to attract them.
    b) Our model has a good predictive accuracy (ROC-AUC = 0.83) and so it can be deployed in future business decision making.