Summary

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:

1. **Cleaning data:**
The data was treated for missing values. The values 'select' were replaced with null value since it did not give us much information. Columns with more than 40% missing values were dropped, for rest, imputation was done based on their respective count plots. Finally very few rows were left with null values, we dropped those rows.

2. EDA:
We performed univariate analysis for categorical and numerical variables. It was found that a lot of categorical variables were not adding any relevant information to the model, therefore, we dropped them. The numeric variables were good and no outliers were found.

3. **Dummy Variables:**
Dummy variables were created for categorical variables excluding binary variables like 'Do not email', 'Do not call'. The original columns for which dummies were created were dropped later.

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

5. **Scaling:**
StandardScaler was used to scale numerical variables.

6. **Model Building:**
Firstly, RFE was done to attain the top 15 relevant variables. Few variables were dropped later (manually) depending on the p-value and VIF values.

7. **Prediction on train data:** We did prediction on the train data using the final model.

8. **Model Evaluation(on train data):**
A confusion matrix was made. We got accuracy of 78.96% , sensitivity of 73.49% and specificity of 82.41%, Precision of 72.49% and Recall of 73.49%.

9. **Prediction on test data:**
Prediction was done on the test data frame.

10. **Model Evaluation (on test data):** Accuracy of 79.54%, sensitivity of 73.44% and specificity of 82.97%.

11. **Lead Scores** were calculated, **Hot Leads** were identified, **and Feature importance** was identified using **res.params.sort_values(ascending=False).**

12. Recommendation:
The company should make calls to the following leads as these are more likely to get converted:

- The leads with the lead source as "Welingak Website"
- The leads who are "working professionals"
- The leads with origin as 'Lead Add Form'
- The leads who spent "more time on the website"

The company should NOT spend too much time on the following type of leads as these are less likely to get converted:

- The leads with last activity as "Olark Chat conversation"
- The leads whose Specialization was "Others"
- The leads with lead origin as 'Landing page submission'
- The leads whose last notable activity was 'Email opened'
- The leads whose last notable activity was 'Page visited on website'
- The leads with last notable activity as "Olark Chat conversation"
- The leads who chose the option of "Do not Email" as "yes"
- The leads with last notable activity as "Modified"
- The leads with last notable activity as "Email link clicked"

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

....

Submitted on 21-01-2025 by Radhika Mahajan, Aqib Jallal and Aishwarya Girhare