Summery

1. **Data Preparation: Missing Value Handling**

- There are some categorical features having a label as “SELECT”. This means the person might not have selected any value for that field. Hence this is as good as a missing value. So converting SELECT into the NaN
- After identifying all the missing data, dropped columns having more than 70% null values
- As the Lead Quality depends upon the intuition of the employee, it will be safer to update the NaN to “Not Sure”
- There are too many variations in the columns (‘Asymmetrique Activity Index’,’Asymmetrique Activity Score’,’Asymmetrique Profile Index’,’Asymmetrique Profile Score’) and it is not safer to impute any values in the columns and hence we will drop these columns with very high percentage of missing data
- We can impute the MUMBAI into all the NULLs as most of the values belong to MUMBAI
- Since there is no significant difference among top 3 specialization , hence it will be safer to impute NaN with Others
- For Tags column, more than 30% data is for “Will revert after reading the email” and hence we can impute NULLS with Will revert after reading the email
- More than 99% data is of “Better Career Prospects” and hence it is safer to impute NULLS with this value
- More than 85% data is of “Unemployed” and hence it is safer to impute NULLS with this value
- More than 95% data is of “India” and hence it is safer to impute NULLS with this value

2. **EDA (univariate analysis, outlier detection, checking data imbalance)**

- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of ‘API’ and ‘Landing Page Submission’ Lead Origins and also increasing the number of leads from ‘Lead Add Form’
- To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of ‘Google’, ‘Olark Chat’, ‘Organic Search’, ‘Direct Traffic’ and also increasing the number of leads from ‘Reference’ and ‘Welingak Website’
- Websites can be made more appealing so as to increase the time of the Users on websites
- We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent

- To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads
- We also observed that there are multiple columns which contains data of a single value only. As these columns do not contribute towards any inference, we can remove them from further analysis

3. Dummy Variable Creation

As logistic regression can worsk with numeric data only, creating dummy variables for the categorical columns.

4. Splitting Data into Training and Test set

Next, the dataset was split into training and test set, to train model first with a chunk of data and then evaluate its performance on unseen data.

5. Feature Scaling

- Feature Scaling is required before Logistic Regression to bring all the features in same scale; this ensures that features with high magnitude are not given higher importance by Logistic Regression Model.
- Currently the company has 37.85% conversion rate. This means among the targeted people only 37% are converting into customers.

6. Model Building

(Feature Selection Using RFE, Improvising the model further inspecting adjusted R-squared, VIF and p-vales)

All variables have a good value of VIF. But we observed earlier that the column "Tags_invalid number" has high p-value and hence we will drop this column and remake the model.

7. Final Model and Model Evaluation

Now that we have the final set of features obtained by removing highly collinear ones, using RFE, inspecting p-values and VIF- we can build the final logistic regression model and evaluate its performance

8. Conclusion

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on their website since this helps in higher conversion
- Increase on sending SMS notifications since this helps in higher conversion
- Get TotalVisits increased by advertising etc. since this helps in higher conversion
- Improve the Olark Chat service since this is affecting the conversion negatively