Summary

1. First we have to inspect the dataset, it has 9240-No of Rows and 37- No of columns.
2. Check for the Statistical aspects of the dataframe
3. Checking for missing values (column-wise) and the percentage of missing values.
4. Tags , Last Activity and Last Notable Activity contain remarks from the sales executive who make calls to the potential customers. We can't use this data generated by sales team for the purpose of building a model that calculates lead score. So these columns can be dropped.
5. Dropping the columns with 40% or more missing values
6. Imputing missing values in columns with high number of null values¶
7. The distribution of the data is heavily skewed. Therefore we can drop such columns.
8. Numerical Variables are Normalised
9. Dummy Variables are created for object type variables
10. Looking at correlations and Dropping highly correlated dummy variables
11. Total Rows for Analysis: and Total Columns for Analysis:
12. Splitting the Data into Training and Testing Sets
13. The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
14. Use RFE for Feature Selection and Running RFE with 10 variables as output.
15. Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
16. Predictions on test data set and Overall accuracy 71.4%
17. The ROC Curve should be a value close to 1. We are getting a good value of 0.80 indicating a good predictive model.
18. From the curve above, 0.3 is the optimum point to take it as a cutoff probability.
19. **The lead source are**: Google , Olark chat ,Organic search ,Reference

Conclusion and Suggestions:

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, X company needs to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

First, they should sort out the best prospects from the leads generated. 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' contribute a lot towards the probability of a lead getting converted.

They should monitor each lead carefully so that the information and communication sent to leads is tailor made for that type of lead. Course offerings and information should be customized in a way that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.

They should focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.