Summary Report:

The problem statement was to come up with a model that can correctly identify the "*hot leads*", i.e customers who would be interested in enrolling for the courses offered on the platform of X education. The initial dataset given contained 37 variables. This data was treated for null values and outliers. The columns were also filtered based on the variance present in them. A significant number of columns had to be dropped on account of negligible variance.

Some columns such as *"Specialization"* and *"Tags"* were grouped into broader buckets so that the data becomes more readable with less levels for categorical columns.

Exploratory data analysis was done after the treatment of the data and the following observations were made.

- Converted leads had a *Medim* level of activity index whereas the non-converted leads tend to be in the *High* activity category.
- Total visits on the website surprisingly had no effect on the lead conversion rate.
- Time spent on the website was comparatively higher for the leads that were converted.

The data was then made ready for building a model on and therefore the categorical data was converted into dummy variables and binary variables (for only 2 categories). The data was further scaled (standardised) so that the interpretation of the model obtained becomes clearer. The final tally of the number of columns was 69.

The following methods were applied to build the model.

Principal Component Analysis:

Principal components were obtained on the training dataset (70% of the initial data) and a scree plot was made, based on which 20 principal components were chosen out of 69 since these 20 components were able to explain around 95% of the variance.

A Logistic regression model was built on these 20 components and the accuracy of the model so obtained on the test data came out to be 91%.

Although PCA is really fast and handy, it fails to explain the variables that have a role in increasing or decreasing the lead conversion rate.

Recursive Feature Elimination:

RFE was used on the data with 15 variables and a logistic regression model was built on these 15 variables. It was found that one variable was insignificant and therefore was removed from the model. The model at last contained 14 variables with p-values less than 0.05 and a variance inflation factor (VIF) below 3. The model so obtained was stable with no multi-collinearity.

The lead score was assigned to each of the leads and the optimal cut-off point was chosen by plotting the sensitivity, specificity and accuracy of the model for all points. The optimal cut-off was found to be 41%, i.e a probability of more than 41% would mean that the lead would be treated as converted.

The model was tested on the test data and the accuracy was found to be 86% which means that the model is correctly able to classify around 86% of the points in the dataset. The values for sensitivity and specificity were also 83% and 87% respectively.

The accuracy so obtained is less than that obtained via PCA, however the following variables were found to be the drivers of a high lead conversion rate:

- Lead Origin (Lead Add Form)
- Last Activity (Had a phone conversation)
- Last Activity (Form submitted on the website)