SUMMARY

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. Provided data includes a lot of information about the leads time they spent, the source, etc. We have done the following analysis on the given data:

1. Cleaning data: Data given is decent. As to the cleaning of it, we have replaced the value 'Select' with nan because wherever select is the value that means the value is not filled by the lead. Apart from that we have made a small change to the values for columns lead source, last activity. Instead of having so many values for a column when there is less occurrence of a value we replaced it with others.
2. EDA: Exploratory data analysis is done at length both univariate and bi variate analysis has given us good insights. Here is low variation in Page Views Per Visit and TotalVisits but higher variation in Total Time Spent on Website. There are a lot of outliers in Page Views Per Visit and TotalVisits. There is positive correlation between Total Time Spent on Website and Conversion. We also found in our analysis that lot of categorical features doesn't add any value to our resolution.
3. Outlier Analysis and Null value treatment: We found that lot of features has null values and we removed the features with >35% of null values in them. We have also dropped rows with null values in them to get a clean dataset. Outlier analysis is performed later and we removed the outliers accordingly.
4. Dummy Variables: Dummy variables were created after dropping the unnecessary features.
5. Model Building: We have split the training and test data to 70% and 30% respectively from the cleaned dataset. We have used standard scaler to bring the features on to same scale. We have then used RFE(Recursive Feature Elimination) to limit the number features used in the model to 15. We have then built a model using those 15 features and an added constant. Using VIF (Variance Inflation Factor) and the P value we have eliminated the features that are not significant to the model performance. Using ROC curve we have found optimal point to be 0.35 and got the accuracy of 0.79 and 0.79 sensitivity and specificity.
6. Model Evaluation: With the model we have we got an accuracy of 0.78 and sensitivity of 0.79 and specificity of 0.77
7. Precision – Recall: Using precision, recall we found that 0.41 is used as optimal cut off point. We got 0.79 accuracy, 0.82 specificity and 0.74 sensitivity.

   We found that the top features to look out for are:
   - Lead Source
   - Last notable activity
   - Last activity
   - Total time spent on website

   Keeping their efforts focused on the features would help X education to convert the potential buyers to buyers of the course to generate high profits.
