SUMMARY

The Business problem of X Education is solved based on the Data(Lead.csv). The Analysis focused on greater conversion rate based on the leads(Converted Column) given in the problem. The data gave a lot of insights regarding the unemployed, working Professionals through EDA and by analysis we tried to work on certain parameters to get more people to join the courses and ultimately enhance the conversion rate.

STEPS:

1. **Cleaning Data:** Data is partially clean and missing values are managed well and unnecessary column are dropped off (No Use in analysis Part). For few Null values 'Select' option get converted to 'NaN'. Few of Null values changed to 'Not Provided'. After making dummies those columns also got dropped off for model Building.

2. **EDA**: EDA part is smoothly done to get deeper insights of the data (categorical and Numerical) with the target variable (Converted).Numerical columns donot have outliers and some Categorical Columns were irrelevant.

3.**<u>Dummy Variables</u>**: Dummy variables were created for the categorical variables and   for scaling for numerical variables , MinMax Scaler is used.

4.**<u>Train-Test split</u>**: Train-Test split was done 70% and 30% for train and test data respectively.

5. **<u>Model Building</u>**: Model Building was done using RFE with taking 15 variables at once and depending upon the P-value as well as VIF few columns were dropped off to get to the threshold limit P-value and VIF and final model is prepared it. Thresholds of P-value and VIF are:

VIF<5,P-value<0.05.

6. **<u>Model Evaluation</u>**: Further more a Confusion matrix is created and after that ROC Curve to get the optimum cutoff values of sensitivity, accuracy and specificity which came out to be 80% for each.

7. **<u>Prediction</u>**: Prediction is done on the test set data frame in which the optimum cutoff came to be around 0.35 and accuracy, sensitivity and specificity to be 80% each.

8.**<u>Precision-Recall</u>**: Precision-Recall is also done to keep a check on the data with the cutoff of 0.41 and Precision is 73% and Recall is 76% on test data set.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1. Total number of visits.
2. The total time spend on the Website.
3. When the lead origin is Lead add form.
4. When their current occupation are:
   a. working professional
   b. unemployed
   c. student
5. When the lead source was:
   a. Olark chat
   b. Wellingak website
6. When the last activity was:
   a. SMS sent
   b. Olark chat conversation
7. When the last notable activity is unreachable. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.