Summary

Problem Statement:
X Education gets a lot of leads, the typical lead conversion rate at is around 30% which is very poor. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Summary:

Step 1 : Importing Libraries and Data
Step 2 : Data Inspection
Step 3: Data Cleaning:
● Columns with >40% nulls were dropped. Value counts within categorical columns were checked to decide appropriate action.
● Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.
Step 4: EDA:
● Data imbalance checked- only 38.5% leads converted.
● Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight on effect on target variable.
● Time spend on website shows positive impact on lead conversion.
Step 5: Data Preparation:
● Created dummy features (one-hot encoded) for categorical variables
● Splitting Train & Test Sets: 70:30 ratio
● Feature Scaling using Standardization
● Dropped few columns, they were highly correlated with each other
Step 6: Test-Train Split
Step 7: Feature Scaling
Step 8: Model Building:
● Used RFE to reduce variables from 48 to 15. This will make dataframe more manageable.
● Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
● Total 2 models were built before reaching final Model 3 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.

● model 3 was selected as final model with 12 variables, we used it for making prediction on train and test set.

Step 9: Model Evaluation:

● Confusion matrix was made and cut off point of 0.345 was selected based on accuracy, sensitivity and specificity
plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less
performance metrics around 75%.

● we will choose sensitivity-specificity view for our optimal cut-off for final predictions since metrics dropped when we took precision-recall view.

● Lead score was assigned to train data using 0.345 as cut off.

Step 10: Making Predictions on Test Data:

● Top 3 features that contributing positively to predicting hot leads in the model are:

- Lead Source_Welingak Website
- Lead Source_Reference
- Current_occupation_Working Professional

Recommendations:

● Engage working professionals with tailored messaging.

● More budget/spend can be done on Welingak Website in terms of advertising, etc.

● Incentives/discounts for providing reference that convert to lead, encourage to provide more references.

● Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.