Summary

An education company named X Education gets a lot of leads, its lead conversion rate is very poor at around 30%. 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 higher conversion chance. CEO's target for lead conversion rate is around 80%.

Data Cleaning:

1. Columns with >40% nulls were dropped. Value counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, created new category (others), impute high frequency value, drop columns that don't add any value.
2. Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.
3. Other activities like outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.

EDA:

1. Data imbalance checked- only 38.5% leads converted.
2. Performed univariate and bivariate analysis for categorical and numerical variables.
3. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight on effect on target variable.
4. Time spend on website shows positive impact on lead conversion.

Data Preparation:

1. Created dummy features (pd.get_dummies) for categorical variables
2. Splitting Train & Test Sets: 70:30 ratio
3. Feature Scaling using StandardScaler
4. Dropped few columns, they were highly correlated with each other Model Building:
5. Used VIF to reduce variables from 34 to 18. This will make data frame more manageable.
6. Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
7. Total 5 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 4.
8. reg5 was selected as final model with 17 variables, we used it for making prediction on train and test set.

Model Evaluation:

1. Confusion matrix was made and cut off point of 0.35 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 71%,80%.

As to solve business problem CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, we will choose sensitivity-specificity view for our optimal cut-off for final predictions

1. Lead score was assigned to train data using 0.35 as cut off.
2. Making Predictions on Test Data:
3. Making Predictions on Test: Scaling and predicting using final model.
4. Evaluation metrics for train & test are very close to around 80%.
5. Lead score was assigned
6. Top 3 features are:
    a. Lead Source_Google
    b. Lead Source_Olark Chat
    c. Lead Source_Other Social Sites

Professional Recommendations:

1. More budget/spend can be done on Google Website in terms of advertising, etc.
2. Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
3. Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.