Lead Score Case Study- Summary

The below analysis is performed for X Education which will tell us that which industry professionals will join their courses. The given dataset is having lot of information about the customers who visit the site, the time they spend over there, then how they reached the site and the conversion rate. The following technical steps are used:

1. Data Cleaning:

- ✓ Checked unique values of each category column and dropped all those columns having only one unique value.
- ✓ Replaced 'Select' with a null value since it did not give us much information.
- ✓ Dropped null values having percentage higher than 40
- ✓ Handled the missing values by making a separate category of it.
- ✓ Combined some categorical values with same meaning.
- ✓ Treated Outliers using capping 3 Standard Deviation or trimmed the rows.

2. Exploratory Data Analysis:

- ✓ Found some of thecategorical variables were irrelevant. Dropped them.
- ✓ Performed Univariate Analysis for both Continuous and Categorical variables.
- ✓ Performed Bivariate Analysis with respect to Target variable
- ✓ Performed Multivariate Analysis by plotting correlation Heatmap and Scattered Pair Plot.

3. Data Preparation for Modeling:

- ✓ The dummy variables are created for all the categorical columns.
- ✓ Used Standard scalar to scale the data for Continuous variables.
- ✓ The Split was done at 70% and 30% for train and tests the data respectively.
- ✓ Dropped Dummy variables whose information was not precisely given.

4. Model Building:

- ✓ Using Recursive Feature Elimination, we selected 15 features.
- ✓ While building models, we eliminated 3 features based on high p-values.
- ✓ Our final model has 12 features.
- ✓ All features in the model have VIF values of less than 2.
- ✓ All 12 features of our final model are significant.

5. Model Evaluation:

- ✓ After considering, Precision Recall Curve and Optimal cut-off curve, we come up with an optimal cut off rate of 0.39.
- ✓ Confusion Matrix and Model Scores.
- ✓ Cross Validation Scores were also same.
- ✓ ROC Value comes out to be same for both train and test. ROC of 96%.
- ✓ Feature Importance Curve.
- ✓ Findings and Recommendation

	Train Scores	Test Scores	Cross Validation Scores
Accuracy	91.52	91.83	91.38
Specificity	95.78	95.94	95.86
Recall	84.87	85.00	84.39
Precision	92.82	92.62	92.90
Number of Features	13.00	13.00	13.00