X Education Analysis Summary

The entire analysis/case study is based on the problem statement given by X Education, an educational platform for industry professional. The marketing of these courses is done on several websites including Google Search Engine. The data that was provided had information regarding the customer demographic details & their latest activities in the X Education platform, the source & the origin of the lead, the time and the frequency of visits made by the customers, etc.,

X Education had a typical sale-conversion ratio of around 30% which was plummeting and the business wants to increase this going forward.

Steps carried out in the Analysis Process:

1. **Data Preparation/Cleaning:** The data that was provided was having a mix of numerical and categorical variables. There were some variables with more than 45% of missing values. So, we decided to drop those variables. Out of the left variables some categorical variables were filled with Select entries as such and also there were many entries with very low frequency which added no information. So, we grouped them into single category as Other and proceeded with the process.
2. **Exploratory Data Analysis:** Univariate and Bivariate analysis was conducted on the dataset. Univariate Analysis of Numerical variables did show some outliers which were removed. Though the data had a lot of variables as independent variables while doing a multivariate analysis using correlation matrix – heat map with the target variable not all variables was adding value to the model because the correlation co-efficient was low.
3. **Encoding of Categorical Variables:**
    i. **Label Encoding:** All the binary class categorical variable was encoded as 0 and 1.
    ii. **Dummy Variable Creation:** All multi-class categorical variables were transformed by creating dummy variables for different levels and finally dropping the “Other” labels manually mapped to avoid noisy data and also avoid dummy variable trap.
4. **Train – Test Split:** A Train Test Split was created for validation of the model on the unseen data. A 75:25 split was used were 25% of the data was taken for validation set.
5. **Scaling & Transformation:** All the numerical variables were scaled with MinMax Scaler to normalize/standardize the numerical variables present in the data.

6. **Model Building:** First model was built with Recursive Feature Elimination by selecting the top 15 features and then subsequent iteration of models were carried out by manual elimination of variables was carried out based on p-values and VIF. Finally, we arrived at the model which had features with p-value < 0.5 & VIF < 5.
7. **Model Evaluation:** A confusion matrix was built with initial threshold value of 0.5 and then ROC-AUC curve was plotted to get optimum threshold value for cut-off probability. We the calculated the Evaluation metrics – Accuracy, Sensitivity, Specificity, Precision and Recall.
8. **Prediction on the Validation/Test Set:** The final model was used on the unseen data and was evaluated on various metrics to check the feasibility & sustainability of the model. The model did give a good train and test metrics.

Key Learnings:

1. Based on Use Case/Problem Statement/Industry Purpose
    a. Sensitivity & Specificity trade-off
    b. Precision-Recall trade-off
2. Making use of ROC-AUC curve to choose the best performing model of all the different models built.
3. Using Accuracy, Sensitivity & Specificity values for different cut-off thresholds to choose the optimum cut-off threshold probability.