LEAD SCORING CASE STUDY SUMMARY REPORT

This case study has been performed for X Education that aims to improve upon it's Lead Conversion Rate by identifying the right set of Leads and focus upon them (Hot Leads). They need a Business Model which is agile and remains stable in production with changing Business Requirements.

MAJOR CHALLENGES FACED :

1. Figure out and **discard** variables that can cause bias and **keep the right ones for model**.
2. **Discard the data generated by Sales team** to prevent overfitting (caused major shortage of unbiased reliable data for analysis).
3. **Not having enough categorical variables** due to above. To ensure we do not miss out further on any variable, decided to run RFE with all the 18 variables.
4. **Keeping Sensitivity and Recall metrics high**- considering one wrong variable was causing a significant dip in the above two due to skewness. Thus, had to discard **Specialization & City** beforehand.

Steps Followed: -

1. Data Inspection and cleaning

a) Presence of multiple nulls in dataset removed by **dropping** variables having more than **40% nulls**.
b) Columns with '**Select**' value labels (same as null) replaced with **Nan** as customer haven't selected any option.
c) Checked all columns for Duplicates (none) and rows with more than 70% nulls (none)

2. Data Transformation and EDA

a) Data Imputation (with **mode** for biased categorical variables)
b) Data spread visualized and insights drawn for every categorical variable.
-Clubbed low occurrence values together for a better analysis (under **'Others'** tag)
c) Correlation evaluated between **numeric variables ,** which came out to be minimal
d) **Outlier analysis and Treatment** performed for numeric variables

3. Feature Selection and Data Preparation (for model-building)

a) Dummy variables created (for categorical variables) and data encoding performed (for binary variables).
b) Labels added for every Select Category and then first record deleted for each.
c) Train-test split (70:30 ratio)

4. Model Building

a) **RFE** applied on all **18 variables** as output, followed by **StatsModel**.
b) Iterated the process till all variables had **p-value < 0.05 and VIF < 5**
c) **Final number of variables→ 10**

5. Model Evaluation

a) Metrics calculation (from confusion-matrix)
b) ROC Curve plotted and AUC score obtained (0.86)
c) Metrics:- Accuracy: 78.17 Sensitivity: 81.75 Specificity: 76
Precision: 67.5 Recall: 81.75
d) **Trade-off** obtained (at around **0.32**) by plotting **Sensitivity**, **Specificity** and **together**

6. Predictions on Test Data

a) Standard Scalar applied on test set to ensure the same set of variables as Training set.
b) Optimum cutoff of 0.32 (obtained from Train data) applied and metrics calculated.

c) **Metrics** obtained are at par with the Train data metrics, ensuring great performance by our model:- Accuracy: 78.11 Sensitivity: 82.95 Specificity: 75

Precision: 67.63 Recall: 82.95

d) Lead scores have been assigned between **0 -100** based on calculating the probability of a lead getting converted as predicted by our final model.

Conclusion

- Top Features responsible in **determining good conversion rate (as per final-model) :**

  1. Lead Origin_Lead Add Form
  2. What is your current occupation_Working Professional
  3. Lead Source_Welingak Website

- All the above metrics show **promising high scores on Test set** and match well with the training set metrics indicating our model is able to correctly predict & classify the Hot leads.

- As per Business terms, model has an ability to adjust and perform well with the company's changing requirements : **Recall > Precision ensures that potential leads will not be missed out.**