SUMMARY REPORT
LEAD SCORING CASE STUDY

NAYANSHI SAHU

NANDISH JANI

Problem Statement

An education company named X Education sells online courses to industry professionals. Dataset is provided for different variables tracking behaviour of both converted and unconverted leads. We have to build a model to meet the CEO's 80% conversion rate expectation.

Steps involved:

Data preparation, cleaning and EDA

Variables with binary data were mapped to 0 and 1 for model building.

Null values were checked and variables with over 40% null were dropped. Remaining variables with null values were imputed with mean,mode or median based on outliers seen in EDA. Null of some variables were imputed with 'Not Specified' for better business sense and to avoid skewness.

Dummy variables were created for categorical variables.

Train Test Split and Feature Scaling

Data split into train and test. Scaling done using MinMaxScaler.

Correlation Analysis

Dummy variables with high correlation among each other were dropped.

Model Building

RFE was used to select the top 20 variables. Models were built after that and variables with insignificant p-values(>0.05) and high VIF(>5) were dropped. Accuracies were also calculated to observe the difference between each model.

Metrics, ROC, Optimal Cutoff evaluation

Based on the confusion matrix, metrics were calculated. ROC and Optimal Cutoff charts were also plotted.

Precision and Recall

Cutoff was again calculated and found at 0.4. Precision of 0.92 and Recall of 0.87 were calculated.

Test set prediction

Test set prediction was done and key metrics compared with train data.

	Train	Test
Accuracy	0.92	0.92
Sensitivity	0.93	0.90
Specificity	0.95	0.93

Top 5 important variables X Education should focus on for better conversion rate:

1. Tags_Lost to EINS -> 5.963161
2. Tags_Closed by Horizzon -> 5.907348
3. Tags_Will revert after reading the email -> 3.989566
4. Total Time Spent on Website -> 3.651781
5. Tags_Already a student -> (-3.068484)

Conclusion:

All the metrics are fairly high with very little variance between train and test performance. This indicates that the model is working well for the dataset and is not overfitted.

X Education can achieve its CEO's target of 80% lead conversion rate with this model.