Summary of Lead scoring case study for X Education

Based on the problem statement the main focus of the case study was to increase the conversion rate to 80% for X education as prior conversion rate was around 30%. The company want to get an insight on those hot leads based on the existing data from different sources collected.

We are expected to use the data to get insights and build a model to get lead scores for each lead to ensure we are targeting the right lead (hot lead) where the chances of conversion are high and will help us achieve our set percentage of conversion. Targeting wrong audience will utilize a lot of effort and time and would not get great results.

To start with the team worked on data cleaning

1) Checking for missing or null values
2) Checking for select word as that also represent a null or Nan value
3) Dropping columns which are not required or adding to the model like unique values column
4) Understanding the variables and its value counts
5) Removing columns that have no variation in data i.e., >99% single value as variables with more percentage of unique values are zero balance variances and are highly unstable.

Data visualisation:

1) Univariate, Bivariate and heat maps for categorical and numeraical columns

Data preparation:

1) Created dummy variable
2) Train and test split of data
3) Scaling the data

Modelling the data – Used logistic regression

1) Started with feature selection - RFE having 20 variables.
2) Build a model using GLM and check summary for P values
3) Checked the VIF – Variance Inflation_factor
4) Drop variable with P>0.05 and VIF >5
5) Rerun the model to get p <0.05 and VIF<5
6) Adding 'Predicted' column on train data where if Converted_Prob > 0.5 then 1 or else 0

Model Evaluation

1) Confusion matrix and Accuracy

	not converted	converted
not converted	3784(TN)	169(FP)
converted	294(FN)	2125(TP)

the second row and first column (294) are the number of customers who have actually 'converted' but the model has predicted them as not-converted.

2) The overall accuracy of the prediction on train data is 92.73 percent
3) Plotted an ROC curve- trade-off between Sensitivity and Specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
4) Finding the optimal cut off point
5) We have used both - an optimal cut off of 0.3 and 0.7 to check for the better conversion rate.

Checks	O.3	0.7
Accuracy	92.37	91.46
Sensitivity	91.65	80.74
Specificity	92.82	98.03
Positive predictive value	88.64	96.16
Negative Predictive Value	94.78	89.27
FPR	7.18	1.97

6) Based on the converted prob- calculated the lead score on train data

Test data Prediction

1) Scaling the test data
2) Running the final model on the test data
3) Predicting on test data
4) Ran the test data with both 0.3 and 0.7

Checks	O.3	0.7
Accuracy	92.71	91.32
Sensitivity	92.32	80.42
Specificity	92.95	98.05

5) It was noted that both the models were able to predict
6) Calculated the Lead Score on test data

Conclusion

Based on the above steps performed and matrix calculated the overall model is doing great work and having conversion rate of above 80% based on matrix Calculated. The final lead score noted can be used by the sales team interns to make calls as they have higher chances of getting converted.