Summary Report of Lead Scoring Case Study

Problem Statement:

X Education sells online courses and wants to improve its lead conversion rate. Currently the conversion is at 30% and the CEO has set a target of 80%.

Business Understanding:

The Company markets on websites and search engines. When people land on website and fill up a form for the course, they are considered as leads by the sales team. Leads are also generated through past referrals. The Sales Team calls, emails these leads but very few get converted.

Objective:

To target this problem, X Education wants to have a model to identify only the hot leads which need to be nurtured and pursued. These hot leads should have higher conversion chance.

Benefits of the model

- Cost savings by avoiding unnecessary phone calls
- Boost the motivation of the sales team

Data Collection:

The leads data has 9240 records with 37 attributes. The attribute "Converted" is the target variable where 1 show converted leads and 0 show not converted leads.

Selection of the Type of Model:

Data perusal shows we can Build a Logistic Regression Model which will predict the probability of a lead getting converted.

Data Cleaning, Visualization and Preparation:

- Missing Values
  - Attributes with level "Select" – Select updated as missing values since the leads have not filled the same.
  - Deletion of columns with more than 40% missing values, redundant, insignificant columns("Tags"),zero variance columns
  - Missing Values have been imputed using Mode or "unknown".
- Significant outliers identified and deleted "TotalVisits".
- Visualization Tools in matplotlib and seaborn used to visualize the data and draw preliminary inferences
- Conversion of categorical variables to binary and dummy variables.

Train-Test Split and Model Building:

- Data split into X and y and further split into training and testing sets of 0.7 and 0.3 size respectively.
- MixMax Scaler used to scale the data so that the model can predict with higher accuracy.
- Since the total variables were high, RFE was used to select the 20 out of the 45 features.
- Model was iterated multiple items by dropping variables one by one till the p values of all variables where less than 0.05.
- Variance Inflation Factor (VIF) was checked to ensure there is no multi-collinearity. All 14 variables in the final model have VIF<5.
- Probability Predictions on the training set

Evaluation of the Model:

- ROC curve shows the model developed was an optimal model
- The optimal cut-off point for probability was arrived at 0.38 which was used in the final model.
- Accuracy – 81.12%
- Sensitivity -77.27% important metric as focus is on conversions
- Specificity -83.44%

Predictions on the testing set:

Evaluation metrics on the test set

- Accuracy – 80.39%
- Sensitivity – 78.19%
- Specificity – 81.75%

Overall, the model is an optimal one having a good sensitivity score on both training and testing set.

Learnings:

- The model developed should be easily adaptable to meet business requirements. The probability thresholds can be changed to suit the business needs. E.g. period where sales interns are recruited and period where target is met.
- Metrics chosen to evaluate the model is also specific to the business objective.