SUMMARY REPORT-LEAD SCORE CASE STUDY

PROBLEM STATEMENT: An education company X Education , selling online courses to industry professionals The company markets its courses through several websites and search engines like Google. When people browse through these websites or fill up forms providing their email address or phone number, they are classified as a lead. The company is facing a problem of low lead conversion rate of 30% which means out of 100 leads contacted in a day, only 30 gets converted.

BUSINESS OBJECTIVES: The objective of case study is to help the company in making its lead identification process more efficient by identifying leads which are most likely to convert into paying customers or in other words hot leads and thus improve its lead conversion rate to 80%. To achieve this, a logistic regression model is to be built to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

KEY STEPS UNDERTAKEN FOR ANALYSIS:

1. Understanding the Problem statement, Business objectives and gaining necessary domain knowledge.
2. Observing the datasets , checking its structure.
3. Data cleaning of application dataset: missing values and outlier handling, checking datatypes.
4. Carried out Exploratory data analysis (EDA) for better understanding of data and gaining useful insights.
5. Data preparation: create dummy features (one-hot encoded), Train-test split, Feature scaling and looking for co-relations between variables.
6. Feature selection using RFE (Recursive feature elimination) algorithm.
7. Building logistic regression model.
8. Determining the optimum cut of value using ROC Curve, sensitivity-specificity curve and precision-recall curve.
9. Making prediction on train dataset and evaluating the model using metrics: accuracy, sensitivity, specificity, precision, recall.

10. Making prediction on test dataset and evaluating the model using metrics mentioned above.
11. Generating the lead score.

Conclusion:

The logistic model created for above problem statement has following results:

Train dataset:
Accuracy of train data: 0.927
Sensitivity of train data: 0.9229
Specificity of train data: 0.9297
False positive rate of Train Data: 0.0703
False negative rate of Train Data: 0.0771
Precision of Train Data: 0.8922
Recall of Train Data: 0.9229

Test Dataset:

Accuracy of test data: 0.921
Sensitivity of test data: 0.9122
Specificity of test data: 0.9264
False positive rate of Test Data: 0.0736
False negative rate of Train Data: 0.0878
Precision of Test Data: 0.8846
Recall of Test Data: 0.9122

The following variables have positive impact on lead conversion rates:

1. Tags Closed by Horizzon, Tags_Lost to EINS, Will revert after reading the email.
2. last activity SMS sent
3. Total time spent on website.
4. Lead Origin - Lead Add Form
5. Leads source Olark chat, Welingak Website.
6. Current Occupation - Working Professional

The following variables have negative impact on lead conversion rates:

1. Tags switched off , already a student, ringing, Interested in other course and other tags.
2. Last notable activity as modified, Olark Chat Conversation
3. Do not email