Summary Report
Lead Scoring Case Study

Problem Statement

Create a machine learning model for an education company, having online platform for their education courses which predicts and assigns a lead score to each lead based on different variables available from historical leads.

Proposed Solution:

1. Building a logistic regression model to assign a lead score between 0 and 100 to each of the leads
2. Which can be used by the company to target potential leads.
3. A higher score would mean that the lead is hot.
4. We are building a model where model conversion rate is above 80%.

Strategy for Analysing:

1. Reading and Analysing Data.

1.The dataset was loaded into the python notebook studied info, length, shape, describe, size of data.

2. Cleaning Data and Performing EDA.

a). Checking for missing Values. Dropping all values whose missing values is more than 40%.

b) Performing EDA doing Univariate, Bivariate and Multivariate Analysis.

3. Data Preparation

Substituting values 'Yes' with 1 and 'No' with 0. and Creating dummy variable.

4. Splitting Data and Creating Training and Testing Models.

The dataset was split into training and testing datasets in a ratio of 7:3

5. Scaling all Features.

Data points were standardised to similar scale using Standard scaler method.

6. Looking for Correlations Between each Variable.

Using leadData.corrs() in heatmap, we can analyse the correlation between variable.

7. Model Building on Train Set:

Creating object of LogisticRegression Model and using GLM we perform fit_tranform on data and calculated the score. We will use RFE method for further model creation.

8.Feature Selection Using RFE :

a). Assessing the model with StatsModels and creating a dataframe with actual converted flag and the predicted probabilities

b). Checking VIFs.

c). Calculate Confusion Matrix

d). Calculate Accuracy. - 0.8923 i,e 89%

e). Calculate Metrics beyond simply accuracy.

i). Sensitivity -- 0.8296 i,e 82%

ii). Specificity -- 0.9310 i,e 93%

iii). False Positive rate -- 0.0689

iv).Positive Predictive Value -- 0.8811

v).Negative Predictive value -- 0.8986

9.Plotting the ROC Curve
As ROC shows the tradeoff between sensitivity and specificity. We got a good ROC Curve.

10.Finding Optimal Cutoff Point -- 0.375 cut off

a).Calculated Lead Score

b).Again calculated the accuracy - 0.8900 i,e 89%

c).Calculate Metrics beyond simply accuracy.

i).Sensitivity -- 0.8666 i,e 86%

ii).Specificity -- 0.9047 i,e 90%

iii).False Positive rate -- 0.095

iv).Positive Predictive Value -- 0.8486

v).Negative Predictive value -- 0.9164

d).Precision And Recall

i).Precision (TP/TP+FP) -- 0.8811

ii).Recall (TP/TP + FN) -- 0.8296

iii).F1-Score -- 0.8546 i,e 85%.

e).Percision and Recall Tradeoff

11.Making Predictions on Test Model:

a). Calculate Accuracy.-- 0.8870 i,e 88%

b). Calculate Metrics beyond simply accuracy.

i).Sensitivity -- 0.8630 i,e 86%

ii).Specificity -- 0.9028 i,e 90%

iii).False Positive rate -- 0.097

iv).Positive Predictive Value -- 0.8528

v). Negative Predictive value -- 0.9098

12. Conclusions and Recommendations for the Company Strategy:

1.The model evaluation steps revealed that the accuracy, precision and recall parameters are acceptable values.

2. The top three variables in your model which contribute most towards the probability of a lead getting converted are:

a. What is your current occupation.

b. Tags

c. Lead Source.

3.In order to increase the time that a user spends on webpage, the company to employ more web developers and UI/UX designers to improve the experience for the user and thereby luring the users to spend more time on the webpage, exploring the contents.

4.The other valuables like Lead Source especially Google, Direct Traffic, Organic Search, Olark Chat, Reference has a very high chance of getting converted. Welingak Website and Reference has an importance influence on the Lead score. It would be a good strategy to increase the marketing budget on these lead sources.