Lead Scoring Case Study Summary

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. This is a logistic regression problem due to predicting the classification of the leads as converted or not, with probability of conversion to be predicted.
Steps followed:

1. Exploratory Data Analysis and Data Cleaning:

- The dataset was loaded into the python notebook and explored statistically and visually to get an idea of outliers, distribution of data, feature redundancies, and so on.
- Correlations between the variables were also identified using heatmap and pairplots.
- The redundant variables were removed.
- Some of the columns had string values“Select”which meant that the user did not select a particular value in the form. This means it is as good as null values.
- Columns with more than 25% null values were removed from the an model building.
- Outliers were removed statistically from the data from two numerical variables to avoid noise in the model building.
- A few rows with missing values were removed. 98% of the dataset was remaining for the data modelling out of 9240 rows in the original dataset provided.

2. Data Preparation:

- Categorical variables were converted into numerical data of 0 or 1 using dummy variables.
- The dataset was split into training and testing datasets in a ratio of 7:3
- Data points were standardised to similar scale using MinMax scaler for numerical features to avoid bias of some variables in higher scales in the model.

3. Model Building:

- Initial model was created by selecting 15 variables using the RFE technique.
- Insignificant variables that were identified were removed to optmise the model for better model. VIFs were also checked to see the multicollinearity between the model features.
- An important step in logistic regression is to find the optimal cutoff for the probability to fit the business needs and to have good accuracy, sensitivity and specificity. We obtained 0.35 as the cutoff from the plot between accuracy, sensitivity and specificity and probability range. 4.Recall and Precision view was then considered to finalise the cutoffs for the prediction in test dataset. The cutoff in the Recall-Precision graph was obtained as 0.42
- Model evaluation was completed by checking the values of all the performance measures namely ROC curve, between accuracy, sensitivity and specificity, Recall and precision. All of them were in acceptable range. 6.Predicted values were calculated using the model. Lead score is 100 multiplied by the log odd which is predicted, to have a value between 0 and 100.

4. Conclusions and Recommendations for the Company Strategy:

- The model evaluation steps revealed that the accuracy, precision and recall parameters are acceptable values. The Recall score is a bit greater than precision score as well. This fits the business needs for the future.
- Top features for good conversion rate:
  a.Total Time spent on webpage
  b.Lead Source_Reference and
  c.Lead Source_Welingak Website
- 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. The other valuables like Lead Source especially Welingak Website and Reference has an importance influence on the leadscore. It would be a good strategy to increase the marketing budget on these lead sources.
- It is also important not to waste a lot of time on some factors that do not contribute much or negatively affect the leadscores.