Lead Scoring Case Study Summary

Problem Statement

Create a machine learning model for X 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:

Data Understanding

- Determining the data format and type
- Profiling the data
- Exploring the data
- Describing the data
- Ensuring the data quality, accuracy, and validity

Data preparation involves:

- Collecting, cleaning, and labelling raw data
- Exploring and visualizing the data
- Data pre-processing, profiling, cleansing, validation, and transformation
- Pulling together data from different internal systems and external sources
- Transforming raw data into a format compatible with data analytics tools
- Analysing the data

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 heat map and pair plots. The redundant variables were removed. Some of the columns had string values“Select”which means that the user did not select a particular value in the form. This means it is as good as null values.

Model Building:

- Splitting into train and test set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high P-Values
- Check VIF value for all the existing columns
- Predict using test set
- Precision and Recall analysis on test predictions

Model Evaluation (Train):

Confusion Matrix:

Predicted	Converted	Not Converted
Actual		
Converted	3456	445
Not Converted	716	1587

Accuracy:	81.28%
Specificity:	88.5%
Sensitivity:	68.91%

Finding the optimal cut-off point through ROC curve

Precision: 78.1%
Recall: 68.91 %

Model Evaluation (Test):

Confusion Matrix:

Predicted	Converted	Not Converted
Actual		
Converted	1151	476
Not Converted	439	593

Accuracy:	65.58%
Specificity:	70.74%
Sensitivity:	57.46%

Conclusion:

The company has to set sales teams to focus on the below features to convert maximum leads.

Lead Origin_Lead Add Form
Total Time Spent on Website
What is your current occupation_Working Professional
Last Activity_Had a Phone Conversation
Lead Source_Welingak Website
TotalVisits
Lead Source_Olark Chat
Last Activity_SMS Sent