Lead Scoring Case Study – Logistic Regression

X Education had a lower lead conversion rate of around 30. Aim of the case study is to find a possible model that will aim to target a conversion rate of 80%

In order for the model to be built, we need to perform the following:

1. Import the data and inspect the given data so we can validate and build the model.
2. Prepared the data for the analysis – Encode the yes / no variables to 1 and 0 so the logistic valuation can be performed.
3. Perform EDA – Univariate and identified the outliers. Identify the Null values, Convert the 'Select' to Nan and identify the outliers to handle them for efficient analysis. Tags column had several tags which are with very low count – ['In confusion whetehr part time or DLP', 'in touch with EINS', 'Diplomo holder (Not Eligible)', 'Approached upfront','Graduation in progress', 'number not provided', 'opp hangup', 'Still Thinking', 'Lost to Others', 'Shall take in the next coming month', 'Lateral student', 'Interested in Next batch', 'Recognition issue (DEC approval)', 'Want to take admission but has financial problems', 'University not recognized']. These were categorized as Other_Tags for easier analysis.
4. Create Dummy Variables.
5. Split the data Test and Train
6. Feature Scaling
7. Identifying the correlation to avoid multicollinearity
8. Build the model using GLM / RFE / VIF / P-value methods
9. Evaluate the model with the accuracy, sensitivity, specificity

The final formula for the logistic regression model is

0.6405 + 5.3762 * Lead Source_Welingak Website + 5.3912 * Tags_Closed by Horizzon + 5.8772 * Tags_Lost to EINS - 5.0579 * Tags_Ringing – 5.4731 * Tags_switched off – 4.2198 * Lead Quality_Worst + 2.8539 * Lead Source_Reference + 2.0564 * Lead Quality_Might be + 1.9304 * Last Notable Activity_SMS Sent + 1.5860 * Last Notable Activity_Unsubscribed – 1.4668 * Specialization_Other_Specialization + 1.1352 * What is your current occupation_Student

With the model which we built, the training data set has:

- Accuracy: 91.92%
- Sensitivity: 92.27%
- Specificity: 91.70%

And the Test Data has:

- Accuracy: 88.02%
- Sensitivity: 88.17%
- Specificity: 87.94%

The overall precision of the model is 80.67%