Summary Report

The problem statement of the assignment was that X Education needs us to assign a Lead Score to effectively target its leads. Target Lead score is 80%.

Process:

Exploratory Data Analysis of the Data Frame

1. Replaced “Select” values with NULL
2. Saw rows with high % of NULL values and dropped them (with the exception of Tags and Lead Quality since we felt those are important rows to enhance our model)
3. Imputed values for numeric rows like TotalVisits and Page Views Per Visit based on MODE
4. Where we saw imputation wasn’t possible due to spread of values (Lead Source and Lead Activity), we deleted the NULL rows
5. We saw visual analysis of all rows to see how the data was distributed. If the data was very strongly hinged towards one value, those rows were eliminated
6. For missing values in Leads Quality, added the values to “Not Sure” and for Tag, added a new Tag called “Untagged” for the missing values

Data Preparation

1. Converted Yes and No value to Boolean [A free copy of Mastering The Interview]
2. Created dummy variables for all the categorical values
3. Dropped the original rows now that we had the dummy variables
4. For the numeric variables – TotalVisits, Total Time Spent on Website and Page Views Per Website – the values were capped at 95th percentile and top lying outliers were removed. Lower outliers were all 0 anyway so they weren’t touched
5. Conversion rate was checked (37%)
6. Variables with high correlation [abs>=0.6] were removed from the data stream

Test-Train Split and Scaling

1. Data was split into Training 70% and Test 30%
2. Data was scaled using z score scaling method

Model Building and Feature Selection

1. Initial logistic regression model was built on 73 columns
2. With RFE we came down to the most critical 15 columns

3. After inspecting p values and VIF – 3 variables were dropped leaving 12 dummy variables and the constant for final model
4. Without knowing optimal cut off, we took it as 0.5 for conversion
5. Calculated the basic metrics – accuracy (91.85%), sensitivity (86%) and specificity (95.5%) for the current model
6. Also plotted the ROC curve for the same (AUC = 0.96)

Computing Optimal Cutoff Point

1. Computed Optimal Cutoff point as 0.3 based on accuracy, sensitivity and specificity spread
2. Recomputed metrics on this cutoff point – accuracy (91.9%), sensitivity (86.13%) and specificity (95.51%)
3. Computed ROC curve as well (AUC = 0.96)
4. Computed Precision (92%) and Recall (86%). Calculating Precision and Recall Tradeoff also pointed at a cutoff around ~0.3

Predictions on the Test Set

1. Used the data available to forecast probability of conversion and then check on the test set
2. Computed metrics – accuracy (93.07%), sensitivity (88.41%) and specificity (95.78%)
3. ROC AUC = 0.96

Calculating Lead Score

1. Merged the Test and Train Data Sets with Unique ID Prospect ID
2. Lead Score was computed as probability x 100
3. 28.55% of all Leads had a Lead Score of 80 or higher and hence should be the first targeted
4. Key features influencing this were
    a. Tags_Closed by Horizzon
    b. Tags_Lost to EINS
    c. Lead Source_Welingak Website
    d. Tags_Will revert after reading the email
    e. Tags_switched off