Summary for Lead Scoring Case study –

1. Aim was to find out which all variables are directly impacting the conversion rate, so after performing data cleaning on original data we selected only few variables –
    i. Many leads are generated from landing page submission, which was through organic search, direct traffic or through google.
    ii. Very less time is spent on website, may be due to its UI/UX, as conversion is higher if person spends most of its time o website.
    iii. Email opening is the activity which is performed the most when online. Also, very apt segment is spending time on websites – mostly un employed
2. Now after, EDA we divided data to train and testing set – mapped correlation shows high number of features were collinear with each other (multi-collinearity issue).
3. We used RFE , VIF and P-value check for elimination/selection of appropriate features-
    i. Tags_Will revert after reading the email
    ii. Lead Quality_Not Sure
    iii. Total Time Spent on Website
    iv. Lead Origin_Lead Add Form
    v. Tags_Ringing
    vi. Last Notable Activity_SMS Sent
    vii. Lead Source_Welingak Website
    viii. Tags_Closed by Horizzon
    ix. Tags_Busy
    x. Do Not Email
    xi. Tags_switched off
    xii. Tags_Lost to EINS
    xiii. Lead Quality_Worst
4. Then we predicted, and evaluated the model on predicted variables by checking stats metrics by performing confusion metrics -
    i. After selecting apt features through VIF/P-value check
        a. Accuracy – 92%
        b. Sensitivity – 87.7%
        c. Specificity – 94.8%
    ii. Are under ROC curve is 96% , which shows good model fit, to evaluate cut-off point we will check metrics –
        a. Accuracy – 92%
        b. Sensitivity – 90.5%
        c. Specificity – 93.5%
5. We also made prediction on the test set using the same variable used after VIF/p-value check –
    i. Fitting the learnt model on test data below are metrics numbers –
        a. Accuracy – 90.7%
        b. Sensitivity – 88.1%
        c. Specificity – 92.27%
    ii. We also check precision and recall on the data –
        a. Accuracy – 90.85%
        b. Sensitivity – 87.67%
        c. Specificity – 87.05%

So, looking at the metrics we can say that all the variables are mostly correctly predicted whether they are converted or not. Also the lead score mapped were highest for the above mentioned feature variables which indicates that the above variables are very helpful in determining whether the Lead will get ‘Converted’ or ‘Not converted’.