Assignment Summary

We had a dataset provided with ~9k records

Data Cleansing

a. Handling Nulls

- We started off by cleansing by deleting any columns with high (40%) nulls and proceeded with data imputation for remaining columns with >1% to 40% null values
- Post that any rows with null values were dropped

b. Cat Variables

- All cat variables were scanned to see their distribution as per lead conversion status
- Dropped any columns which were heavily skewed as Conversion as 0 or 1
- Rest were cleansed to group the columns with low frequencies
- Any comment columns which are irrelevant for model building were also dropped
- Created the dummy variables for all the relevant cat variables
- Dropped the original columns for which dummy var are created

c. Num Variables

- Scanned columns for outliers and limited the outliers to 95%

Model Development

- Split the dataset into 70%:30% split for Train:Test data
- Did the feature selection with RFE an created the model. Below model was selected based on the regression results for further model evaluation

Model Evaluation

- Model was evaluated by looking at confusion matrix of training data
- Model was checked for Sensitivity, Specificity, False positive rate, Positive Predictive Value, Negative Predicted values
- Optimal cut off was taken as 0.35 as per balanced sensitivity and specificity
- As nos were in acceptable range, the model was evaluated vs the test dataset
- Findings
- The output regression is as follows

Conversion = 3.02 x Lead Origin_Lead Add Form + 2.46x Lead Source_Welingak Website + 2.38 x What is your current occupation_Working Professional – 2.06 x Last Activity_Email Bounced + 1.11 x Total Time Spent on Website – 1.03 x Lead Origin_Landing Page Submission + 1.29 x Lead Source_Olark Chat – 1.33 Last Activity_Olark Chat Conversation +

1.80 x Last Activity_Other_Activity + 1.25 x Last Activity_SMS Sent -0.93 x Specialization_Others + 1.22 x What matters most to you in choosing a course_Better Career Prospects – 1.46

- This shows that top 3 activities that can impact lead conversion are Lead Origin_Lead Add Form, Lead Source_Welingak Website, What is your current occupation_Working Professional >> Hence focus on leads that originated from lead add form, souce is welingak website and is a working professional
- The detrimental vars are Last Activity_Email Bounced, Lead Origin_Landing Page Submission and Specialization_Others >> which means that leads where email bounced or lead originated through landing page and where specialization is others – such leads are less likely to convert