Lead Scoring Case Study Summary

Step 1: Reading and inspecting the dataset- There are 9340 rows and 37 columns.

Step 2: Data Cleaning- Converted 'Select' value to null values in the columns. Next, dropped the columns with greater than 40% null values. For categorical features, imputed null values with mode of the column. Found that in some columns the number of null values were greater than the mode, hence dropped those column. For numerical features, imputed the null values using the median of the column. Next, outliers were capped at 99 percentile.

Step 3: Exploratory Data Analysis- Univariate analysis showed that 12 categorical columns had only one value or the count of other category was negligible. Hence dropped these 12 columns. Bivariate analysis showed that total visits had linear relation with page views per visit. This is understandable because calculation of page views per visit is dependent on number of visitors.

Step 4: Creating dummy variables- Went ahead to create dummy variables for the categorical features.

Step 5: Test Train Split- Used the industry standard split of 70-30 ratio for train-test dataset.

Step 6: Feature Scaling- The numerical columns are scaled so that the absolute values do not have adverse impact on the model co-efficients.

Step 7: Feature Selection using RFE and stats model- Using Recursive Feature Elimination, we selected the top 17 features. Now using stats model, we eliminated features with p-value>0.5 and VIF>5. Finally, we reached to 12 significant features. We then created a dataframe with the predicted probabilities and predicted value. We assumed the threshold of 0.5 to calculate the conversion of the lead. Based on these predicted values, we found the accuracy of the model to be 81.55%

Sensitivity and specificity of the model came to be 69% and 88% respectively

Step 8: Plotting ROC Curve – The ROC curve for the model is also decent with area under the curve to be 89%.

Step 9: Finding Optimal Cutoff Point- Now we plotted accuracy, sensitivity and specificity graph for various probabilities to identify the optimal threshold for our model. The graph cut at 0.35

The accuracy, sensitivity and specificity at this threshold is close to 81% for all three metrics.

Step 10: Finding Precision, Recall and F1 score- The precision score is 79% and recall is 69%. Based on precision and recall tradeoff we got a threshold close to 0.42. F1 score came out to be 74%

Step 11: Making predictions on test set- We applied our model on the test dataset and the accuracy, sensitivity and specificity came out to be 81.49%, 81.27% and 81.63% respectively which is very close to the metrics achieved on training dataset.

Step 12: Conclusion- The model has the following important features determining the lead conversion.

1. Total Time Spent on Website
2. Do Not Email
3. Lead Origin - Lead Add Form
4. Lead Source - Olark Chat
5. Last Activity
    - Converted to Lead
    - Had a Phone Conversation
    - Olark Chat Conversation
    - SMS Sent
6. What is your current occupation
    - Not Specified
    - Working Professional
7. Last Notable Activity
    - Modified
    - Unreachable

The accuracy, sensitivity, specificity of the model are all close to 81% which was required. Hence the model looks good.