Lead Scoring Case Study - Summary

We started by understanding the data provided along with the problem areas.

Case of missing information:

During which it was found that there was an unwanted data tagged as “Select” which did not fit into any of the variables provided. Hence, we decided that this should be a case of missing information which would have been defaulted to Select in the dropdown options. These values were converted into NaN.

Handling missing values:

Once the data was complete, we started handling missing values. First by removing variables which has more than 45% of missing values and then imputing the remaining with a mod

- when the mod iteration is at a higher side and marking them as Not Updated
- when mod did not make sense and dropped the imbalanced data where the data was concentrated on one variable “NO”.

Data Preparation:

Data Preparation started with converting categorical variables first by binary variables then with “dummy variables” and dropped the original columns.

Train and Test data set:

Train and Test data set were prepared using Sklearn with a split of 70% and 30% keeping “Converted” as Target or Dependent variable Scaling was done to convert data between 0 to 1 After assessing we found that we had multiple data variables with very bad P value and co-efficiency which were deleted first using RFE method which gave us 15 preferred variables from which we removed Tags marked as “number not provided” and “wrong number given” as their P value was 0.999 and Co-efficiency was more than 0.24 this gave a suitable model with a VIF of less than 1.05 lowering any possibility of multicollinearity. Now comparing our data model output prediction with dependent variable actual values we drafted a Confusion matrix which yielded 92% of accuracy at a cut-off estimation of 0.5% after ROC analysis and plotting Accuracy, Sensitivity and Specificity in different probable cut-off percentages it was found that the optimum cut-off is 0.3. After resetting our cut-off to 0.3 there was a slight increase in prediction values as even our Precision and Recall trade-off was optimum at 0.3.

Final model:

After this when we run the final model on test dataset the accuracy was at 92% with Sensitivity and Specificity at 86% and 95%.

Conclusion:

During the whole exercise we learnt that Google plays a crucial role in getting prospects into conversion as most conversions is happening by Google searches