Lead Scoring Assignment

Steps followed

1. Exploratory Data Analysis
   1. Reading the data
   2. Data Cleaning- There were lot of missing values such as ‘Lead Quality’ variable had 51% missing values but we got to know it was important by doing bivariate analysis since it was impacting target variable(‘Converted’) more, we could have asked client to provide missing data or we can proceed by deleting the column since imputing is not option it will add bias to data. We proceeded by deleting column.
   3. In the same way deleting columns which has more than or equal to 30% of missing data and below that we can delete rows by proceeding this way we are left with these columns
      1. Prospect ID
      2. Lead Number
      3. Lead Origin
      4. Lead Source
      5. Do Not Email
      6. Do Not Call
      7. Converted
      8. TotalVisits
      9. Total Time Spent on Website
      10. Page Views Per Visit
      11. Last Activity
      12. Search
      13. Magazine
      14. Newspaper Article
      15. X Education Forums
      16. Newspaper
      17. Digital Advertisement
      18. Through Recommendations
      19. Receive More Updates About Our Courses
      20. Update me on Supply Chain Content
      21. Get updates on DM Content
      22. I agree to pay the amount through cheque
      23. A free copy of Mastering The Interview
      24. Last Notable Activity
   4. Tried to impute missing values by finding releation between variables using bivariate analysis but couldn’t find any releation
2. Data Preparation:
   1. Creating dummy variables for categorical variables and scaling the numerical variables
   2. Splitting the data to train and test dataset
3. Model Building:
   1. Using logistic regression since its suitable for binary classification
   2. Selecting features using RFE, first selecting 20 features then eliminating one by one by observing p-value which should be less than 5% and VIF less than 5
   3. Finally we got model with 16 features which is significant, free from multicolinearity and overfitting
   4. “Lead Source, Last Activity, Lead Origin” these are top 3 features influencing conversion
   5. “Lead Source\_Welingak Website, Lead Source\_Reference, Last Activity\_Had a Phone Conversation” are the top 3 dummy variables influencing conversion
4. Evaluation of model:
   1. ROC curve is used for model evaluation which we got as 0.87 which is preety good.
   2. Calculating the optimal cutoff using Sensitivity,Specifictity,Accuracy curve which we found to be 0.35
   3. Since Bussiness requirement is finding Sensitivity because CEO has said to get around 80% lead conversion which is nothing but Sensitivity so choosing cutoff which we got by sensitivity,specificity,accuracy plot
5. Making Predictions on the test data set
   1. The variables are scaled first using just transform and predicted probabilities then using same cutoff as 0.35 we predicted Conversion.
   2. We got accuracy=sensitivity=specificity=0.78 which is near to 0.8 which we got from train data set
6. Calculating the Lead scores by multiplying probabilities by 100 and rounding them to 0 decimals, higher the probablites higher the conversion chance and vice versa, so business can decide cutoff and call the customers with high lead scores, we can suggest them cutoff to be 35 for which we are reaching our target of 80% lead conversion