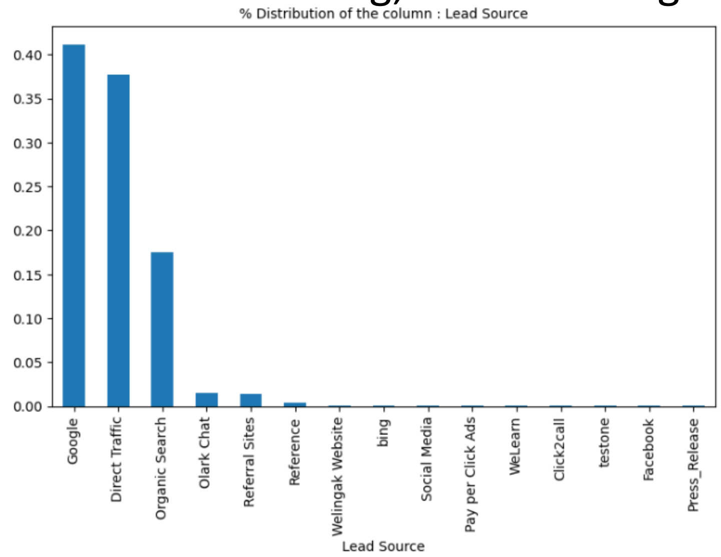
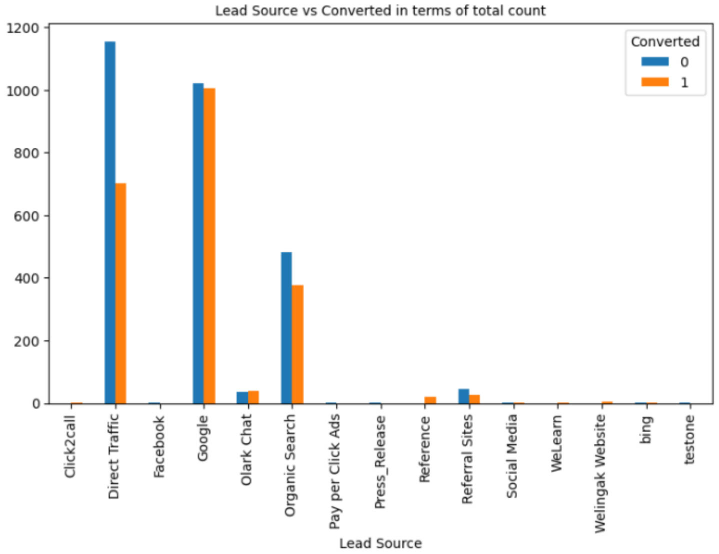
Lead Scoring Case Study

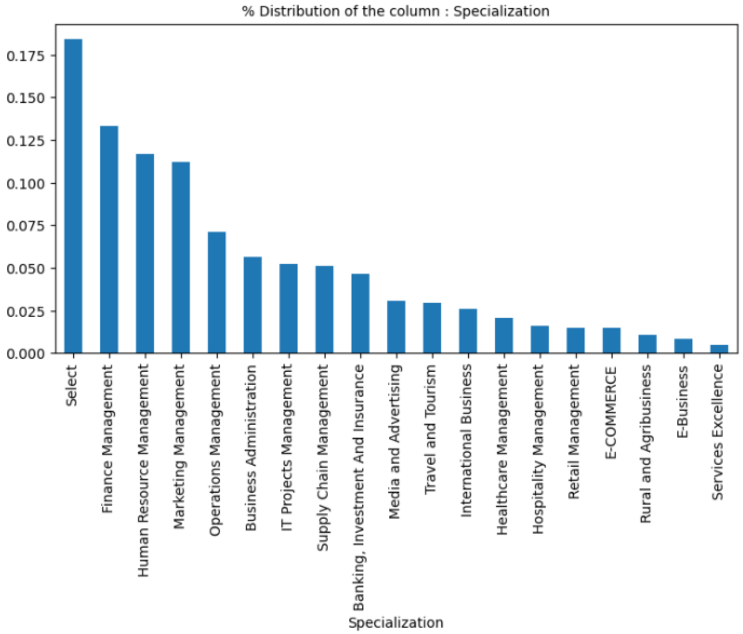
Problem Statement

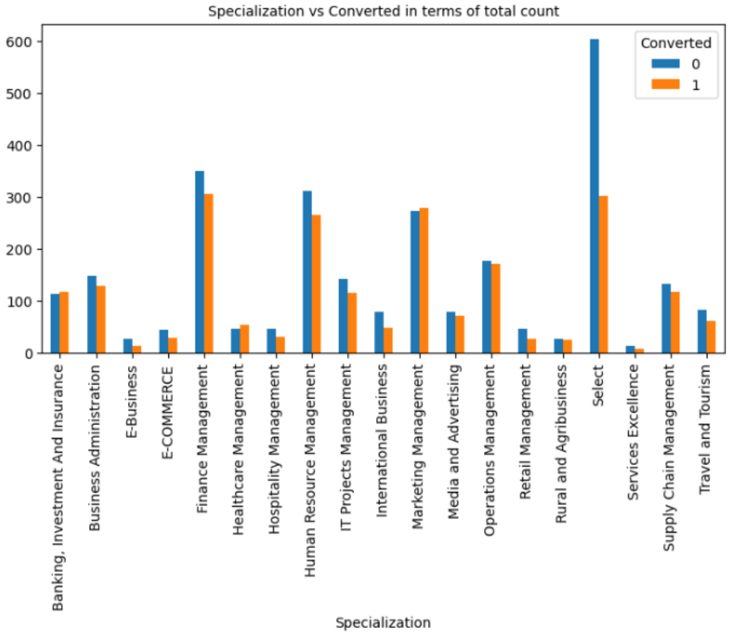
An education company named X Education sells online courses to industry professionals.The company markets its courses on several websites and search engines like Google.When these people fill up a form providing their email address or phone number, they are classified to be a lead.Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.Although X Education gets a lot of leads, its lead conversion rate is very poor.To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

The plots of Categorical variables with respect to Target variable shows that the lead sources like Google,Direct traffic and organic search have good conversion rates. Along with the leads from Marketing,Finance backgrounds have chosen the courses

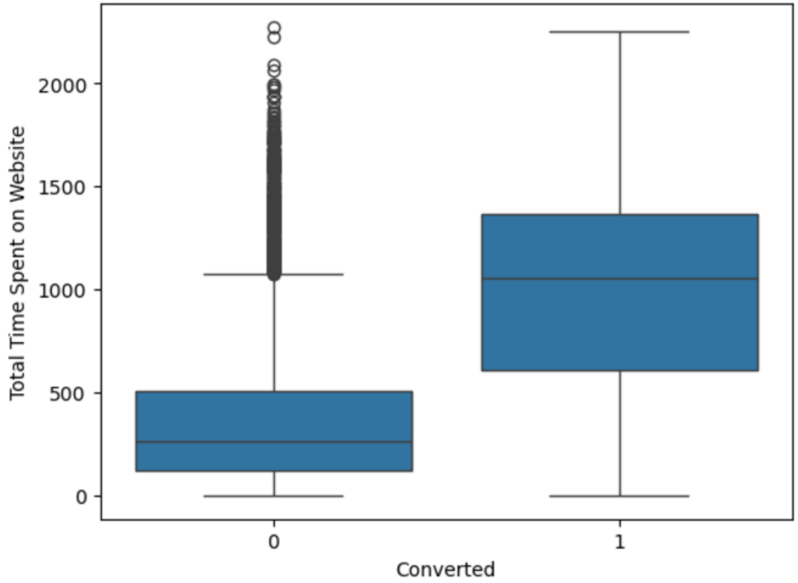


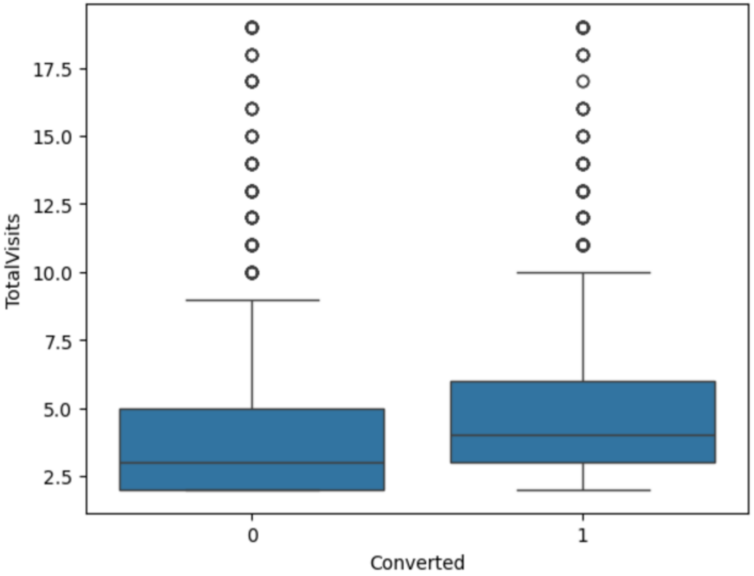


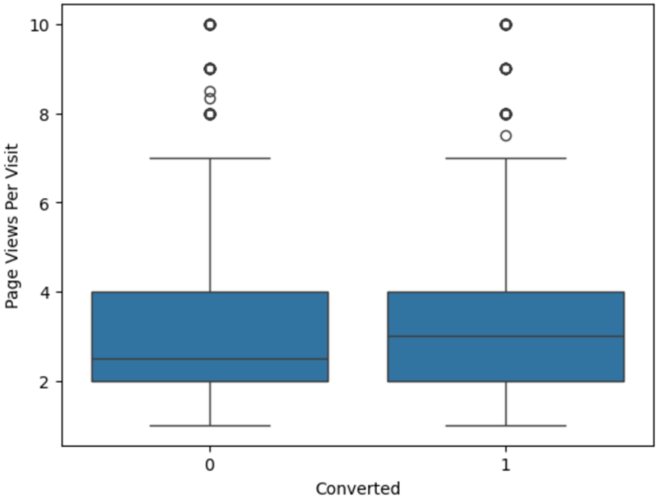




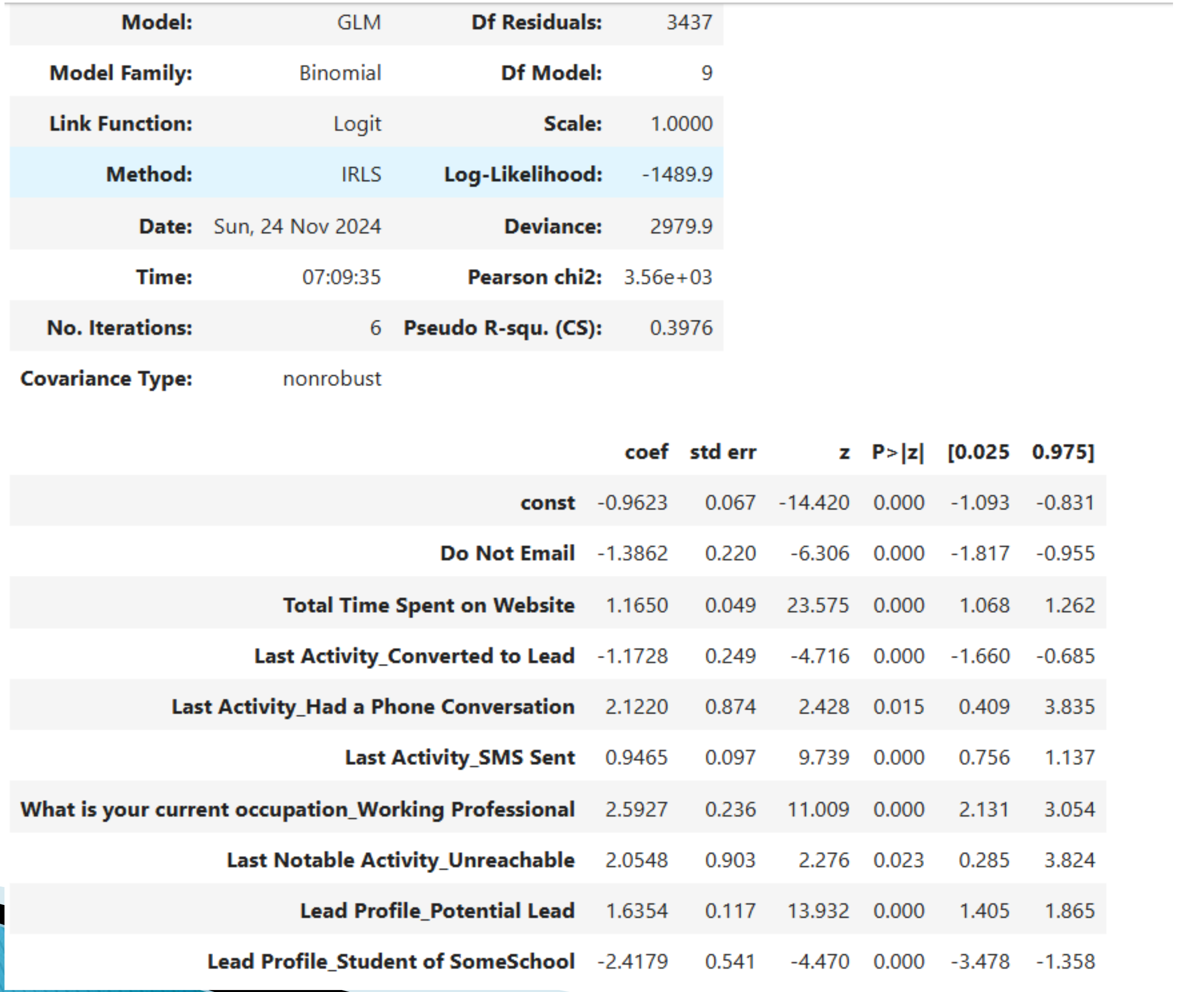
Segmented Univariate Analysis: The plots of numeric variables with respect to the target Converted variable shows that the Total time spent on website has a better lead conversion rate followed by TotalVisits and Page Views per visit.







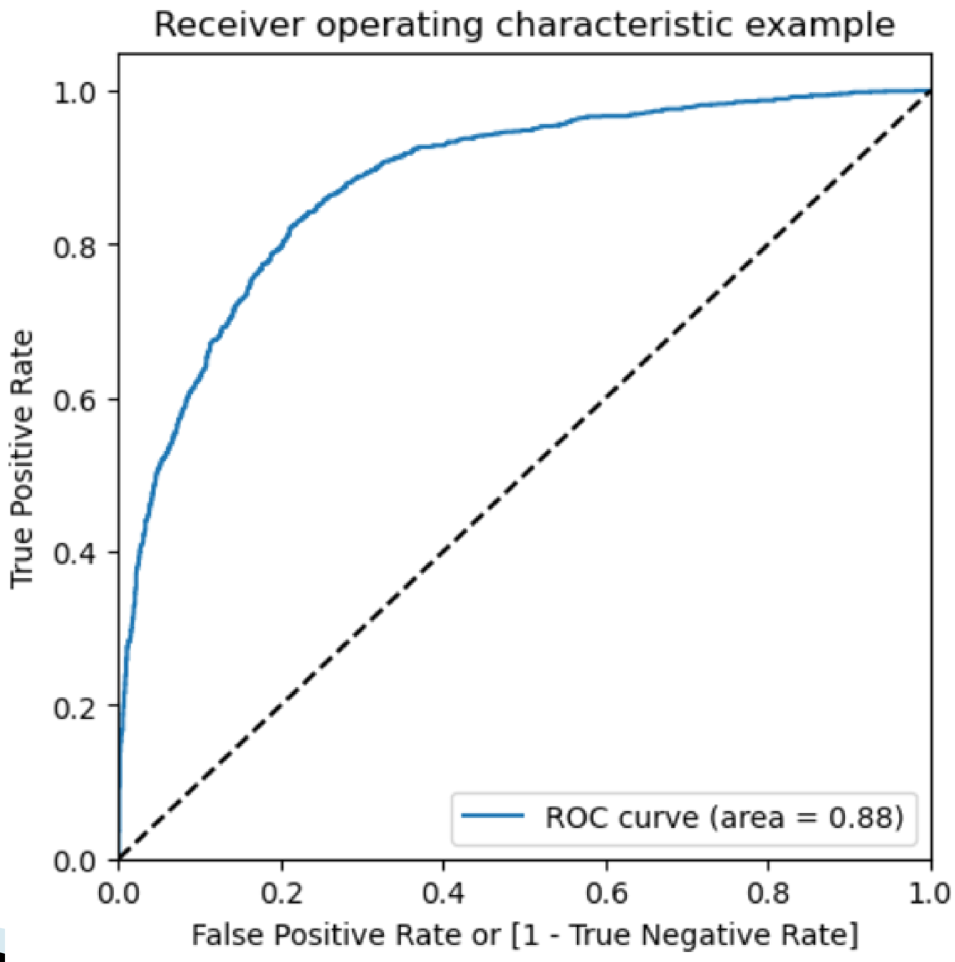
Top three variables in model which contribute most towards the probability of a lead getting converted:-
1)Total Time Spent on Website - Indicating that the more time a lead spends on the website, the higher the likelihood of conversion.
2)Last Activity: Had a phone conversation - Leads that have their "Last Activity" as "Phone conversation" often exhibit a strong positive association with conversion.
3)What is your current occupation: Working professional - Leads who are working professionals are contribute to a higher intent to engage.



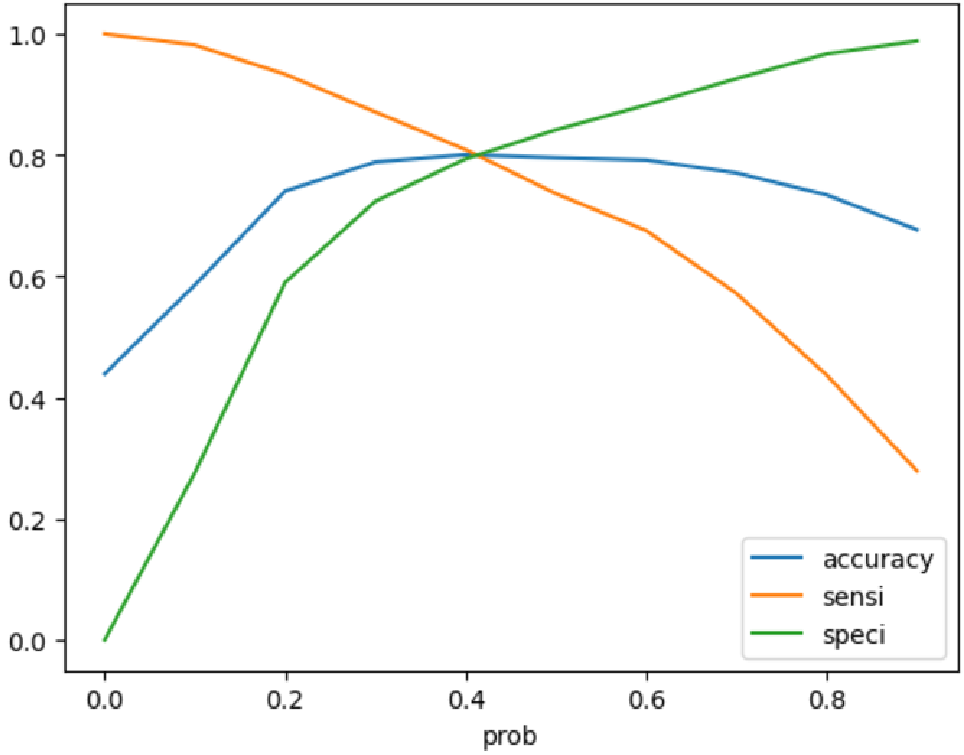
```
          Model:              GLM       Df Residuals:      3437
   Model Family:         Binomial           Df Model:         9
  Link Function:            Logit              Scale:    1.0000
         Method:             IRLS     Log-Likelihood:   -1489.9
           Date: Sun, 24 Nov 2024           Deviance:    2979.9
           Time:         07:09:35      Pearson chi2:   3.56e+03
 No. Iterations:                6  Pseudo R-squ. (CS):   0.3976
Covariance Type:        nonrobust

                                                    coef  std err        z  P>|z|  [0.025  0.975]
                                        const    -0.9623    0.067  -14.420  0.000  -1.093  -0.831
                                 Do Not Email    -1.3862    0.220   -6.306  0.000  -1.817  -0.955
                    Total Time Spent on Website   1.1650    0.049   23.575  0.000   1.068   1.262
                  Last Activity_Converted to Lead -1.1728   0.249   -4.716  0.000  -1.660  -0.685
            Last Activity_Had a Phone Conversation 2.1220   0.874    2.428  0.015   0.409   3.835
                         Last Activity_SMS Sent   0.9465    0.097    9.739  0.000   0.756   1.137
What is your current occupation_Working Professional 2.5927 0.236  11.009  0.000   2.131   3.054
               Last Notable Activity_Unreachable   2.0548    0.903    2.276  0.023   0.285   3.824
                     Lead Profile_Potential Lead   1.6354    0.117   13.932  0.000   1.405   1.865
               Lead Profile_Student of SomeSchool -2.4179   0.541   -4.470  0.000  -3.478  -1.358
```

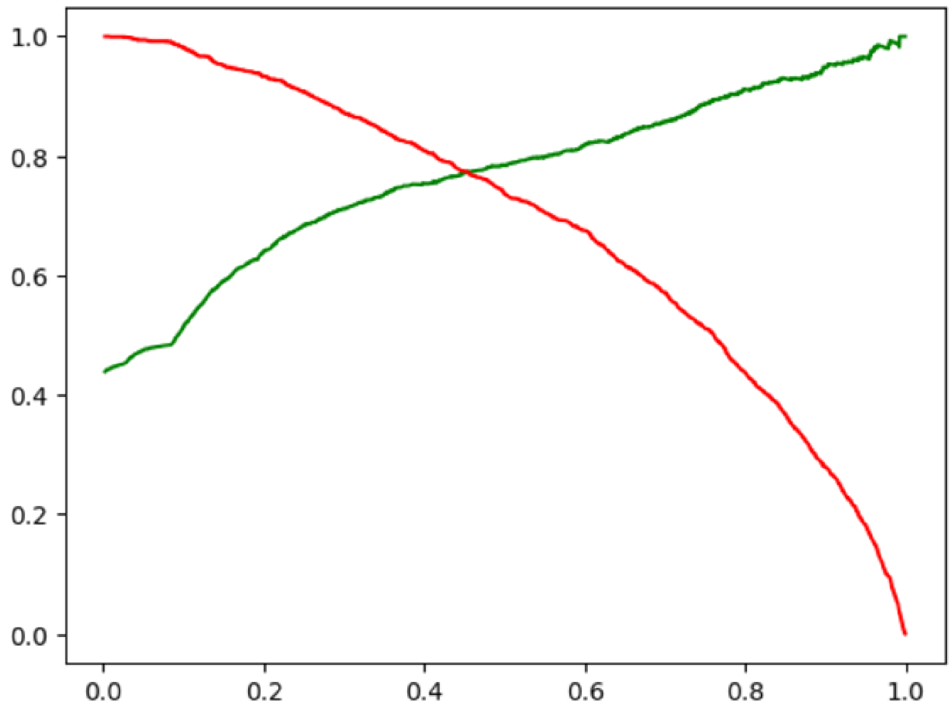
The ROC curve provides the understanding of the classifier's accuracy. The closer to a right angle the curve, the more accurate the model. The classification threshold that returns the upper-left corner of the curve—minimizing the difference between TPR and FPR—is the optimal threshold. The larger the area under the ROC curve better is the model . Here the area is 0.88



The plot between accuracy,sensitivity and specificity shows the variation with various cut-off points.The plot below where the specificity and sensitivity curves cross each other is the optimal cut-off point which is 0.4



Precision signifies how often a ML model is correct in predicting the target variable in this case the Converted variable, the precision comes out to be 76% and recall is 78%. The PR curve which is the trade-off between Precison and recall gives a threshold value of 0.44 as shown in the plot



Steps Followed

1. **Data Cleaning**:

a.The dataset contained missing values, with features like Total Visits and Page Views per Visit were dropped from the dataset.

b.Columns such as Tags and Lead Quality with over 35% missing data were removed.

c.Dummy variables were created for categorical features

2. **Exploratory Data Analysis (EDA)**:

a.Patterns revealed that **68.5% of leads** spending over 3 minutes on the website converted.

b.Categories like Lead Source: Direct Traffic and Last Activity: Phone conversation exhibited strong conversion tendencies.

3. **Model Development**:

a.A logistic regression model was built to predict the likelihood of lead conversion.

b.Recursive Feature Elimination (RFE) was used for feature selection, identifying the top 15 variables.

c.Variables with VIF < 5 and p-value < 0.05 were retained to ensure multicollinearity was addressed.

4. **Model Evaluation**:

Performance metrics included:

Accuracy: 80%

Precision: 76%

Recall: 78%

An optimum cut-off probability of 0.4 was determined using an ROC curve for balanced performance

Thank you