Lead Score Case study
Presentation

By
D G Rani
Sohini Dey

Steps involved

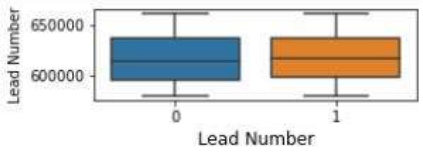
1. **Data Reading and Understanding**
2. **Data handling**
3. **Exploratory data analysis**
4. **Logistic Regression model development**
5. **Making predictions on test data**
6. **Comparing the results with train data**

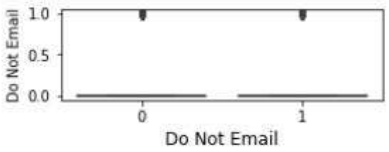
Analysis

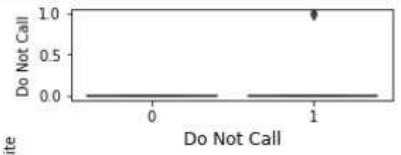
- In the first step we import the libraries and data and take a deep look into the data to understand the variable types means categorical and numeric variables
- In the second step look for missing values. here we dropped the columns with more than 40% of missing data and used imputation technique to replace the missing values

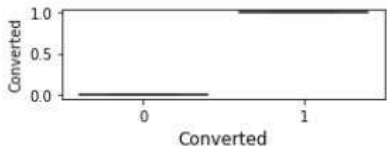
- In the third step we do univariate and bivariate analysis on the data. Here we will get the relation of dependent variable with independent variables. Based on the analysis we drop some columns which are not useful in model building.

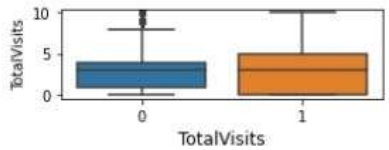
Bivariate analysis for numeric variables

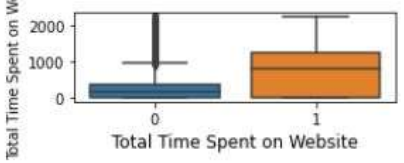


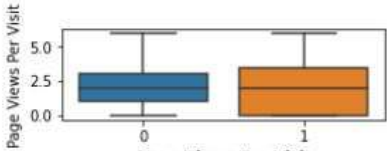


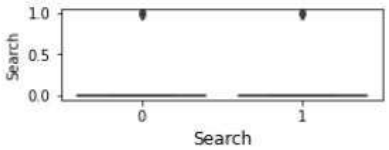


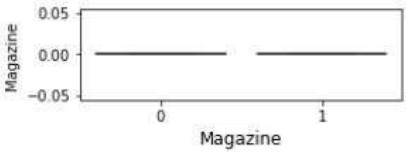


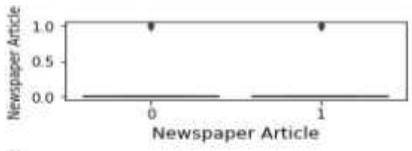


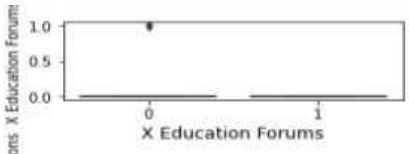


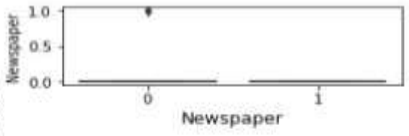


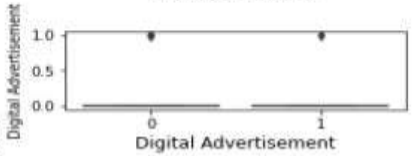


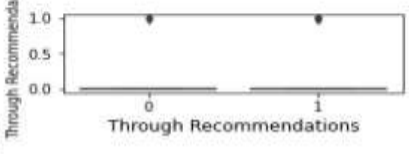


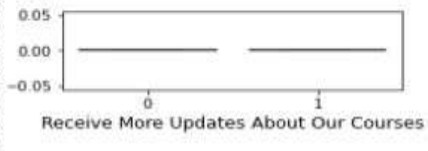


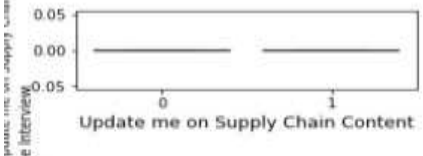


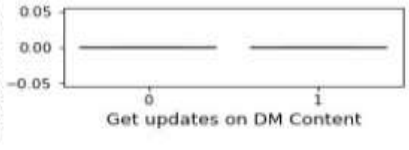


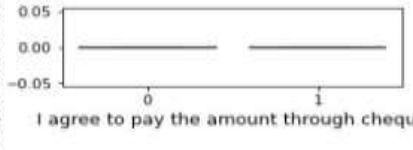


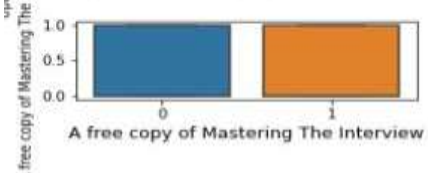




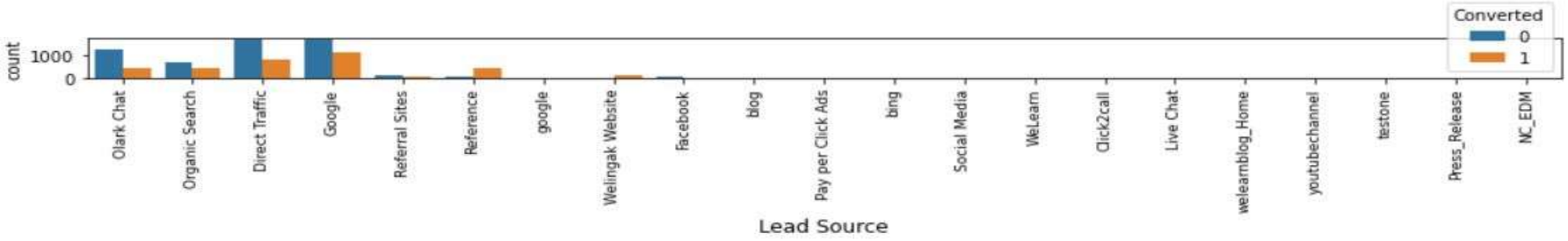


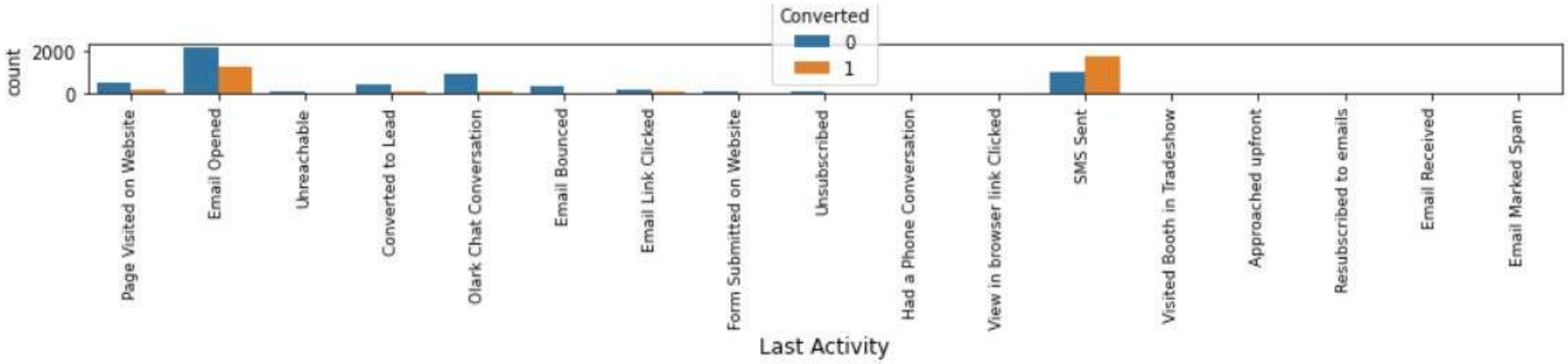




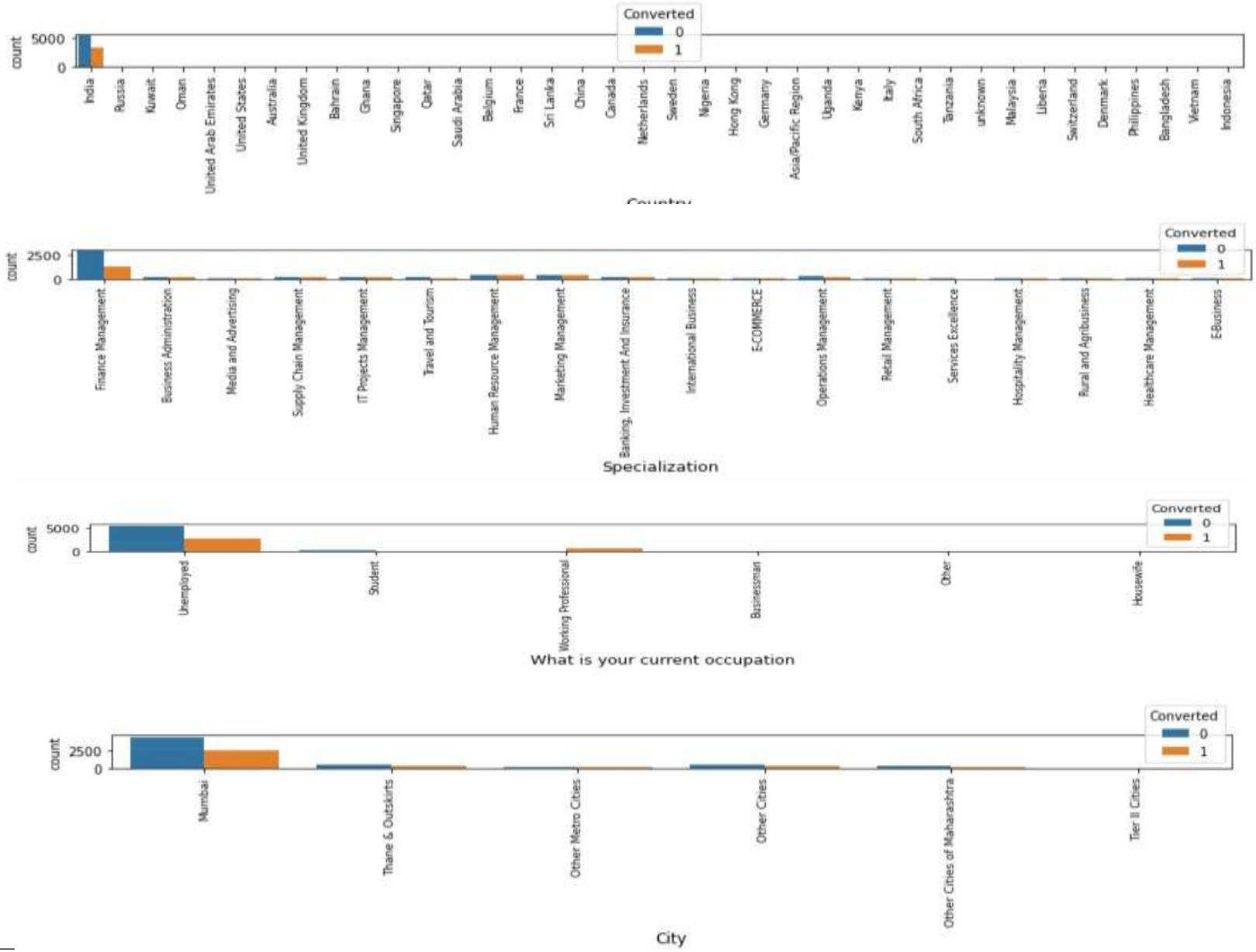


Bivariate analysis of categorical variables

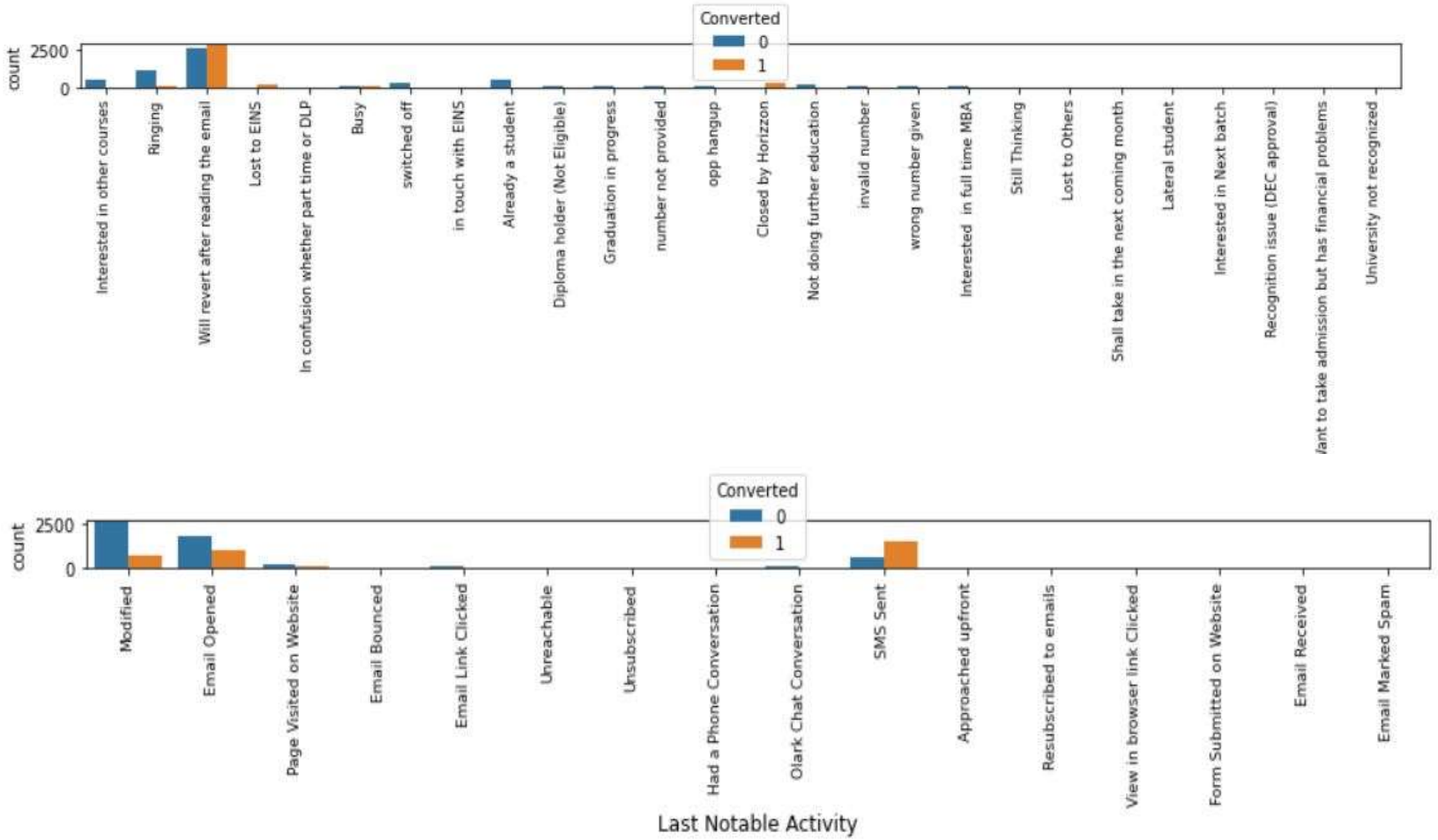




Bivariate analysis of categorical variables



Bivariate analysis of categorical variables



From the above analysis we can say that only some features have high conversions so keep them and drop the other features

- In the forth step we will start model building but before that we need to prepare the data.

  1. first create the dummies to categorical variables

  2. splitting the data into Train and test data

  3. Rescaling the Train Data using Standard scalar

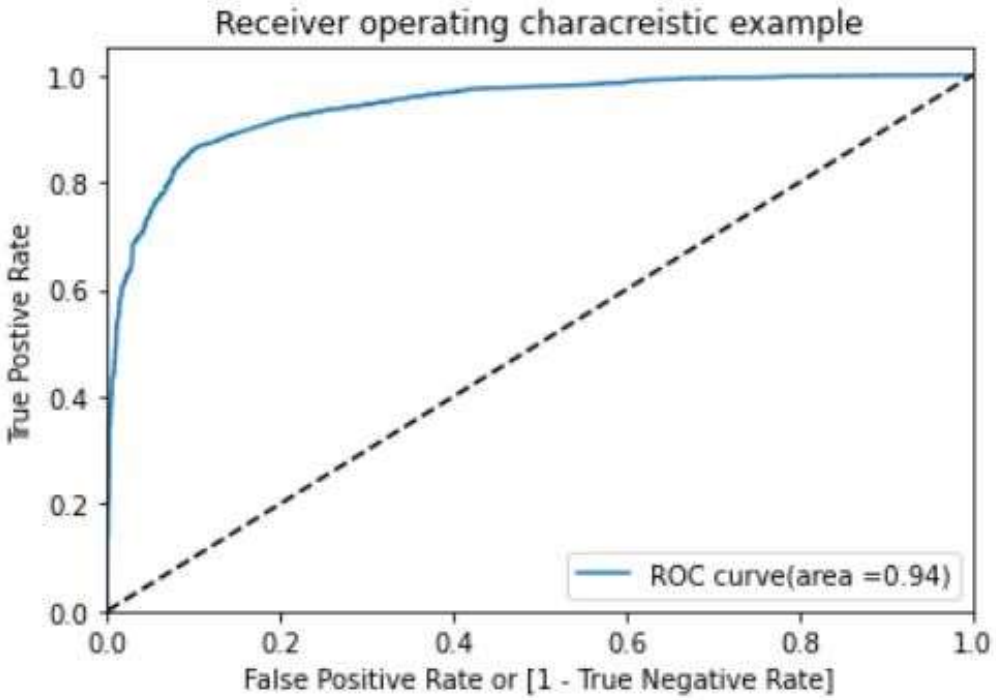
  4. using RFE and states model generate a model and using VIF(<5) and probability eliminate the features to reduce multicollinearity .

  5.after getting low VIF values start to derive the probabilities, Lead Score, Predictions on Train Data

  6. then form confusion matrix to calculate the Accuracy, Sensitivity ,Specificity , Precision.

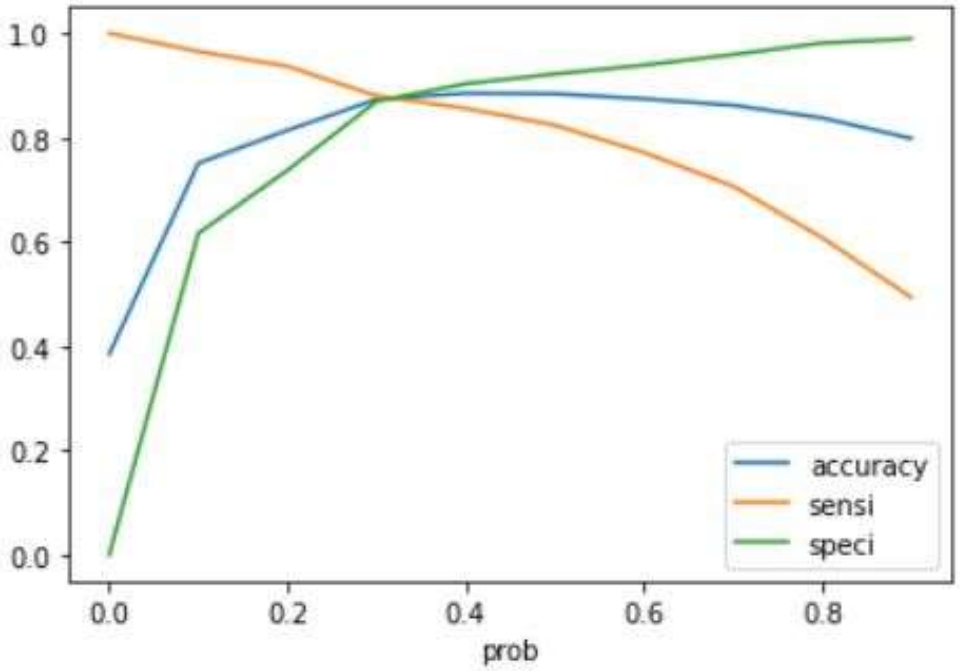
- In the fifth step make the predictions on test data.
- In the last step we compare the Accuracy, Precision , Sensitivity of Train Data and Test Data
- Based on ROC curve Accuracy-Sensitivity-Specificity curve and Precision-Recall tradeoff curve we tell that our model is best Fit

ROC curve



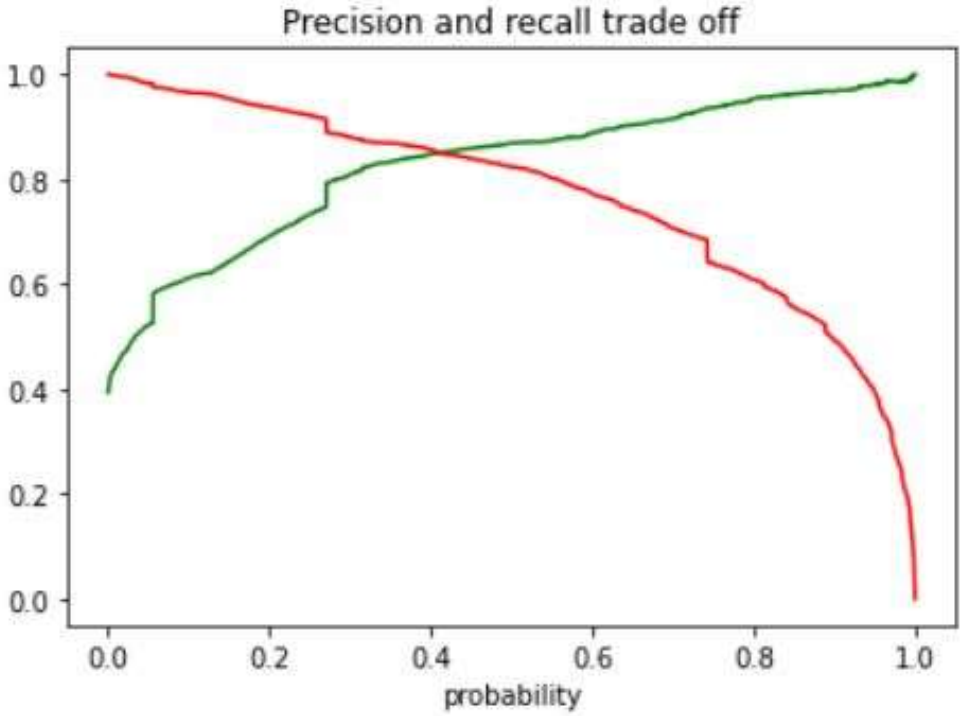
From the ROC Curve value is close to 1, getting a value of 0.94 which is a good predictive model

Accuracy- Sensitivity -Specificity Curve



from the above we can see that 0.3 to 0.4 is our cutoff

Precision – Recall Tradeoff curve



from the precision and Recall Trade-off curve we got cutoff 0.4 which is near to accuracy-sensitivity-specificity curve .so our cutoff is optimum

Comparing Results

For Train Data

- Accuracy = 86.8%
- Sensitivity = 82.3%
- Specificity = 92.1%
- Precision = 80.8%

For test data

- Accuracy = 81.4%
- Sensitivity = 83.4%
- Specificity = 80.2%
- Precision = 70.69%

from the results we can tell that our model is not over fitting the test data

Parameters of final Model

1. const -0.147054
2. Total Time Spent on Website 1.185175
3. Lead Origin_Landing Page Submission -0.708712
4. Lead Origin_Lead Import 1.750642
5. Lead Source_Olark Chat 0.816538
6. Lead Source_Reference 3.278883
7. Lead Source_Welingak Website 4.867717
8. Last Activity_Email Bounced -1.312855
9. Last Activity_SMS Sent 2.049044
10. Last Activity_Unreachable 1.436066
11. Last Activity_other_activity 1.046547
12. Specialization_Finance Management -0.607012
13. What is your current occupation_Working Professional 2.602816
14. Tags_Closed by Horizzon 5.727688
15. Tags_Graduation in progress -1.409163
16. Tags_Interested in full time MBA -2.962110
17. Tags_Lost to EINS 5.472026
18. Tags_Not doing further education -4.144327
19. Tags_Ringing -4.692220
20. Tags_other_tags -3.022055
21. Tags_switched off -4.388958
22. Last Notable Activity_Modified -1.832883
23. Last Notable Activity_Olark Chat Conversation -1.830995

Conclusion

1.From the model we can tell that some coefficients positively related and some coefficients negatively related.
2.Tags_Lost to EINS,Tags_Closed by Horizzon these to are strongly related to the conversion probability but these are dummy variables from tag's categorical column. so “Tag's” can be the important feature to predict conversions. Similarly “ Lead source” and “What is your current occupation” are also important features to increase the converting rate.
3.Final model has Sensitivity of 0.83 , this means it predicts 83% customers out of all the converted correctly
4. Final model has Precision of 0.7069, this means 70.69% of predicted hot leads are True Hot Leads.
5. X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted. for this If we reduce the cutoff to minimum means cutoff > 0.1 are hotleads we will get more hot leads
6. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company’s aim is to not make phone calls unless it’s extremely necessary, i.e. they want to minimize the rate of useless phone calls.For this we need to use the maximun cutoff for promising hotleads means cutoff>0.9 are hot leads.