Lead Score Case Study

Group members
1-Saba gul
2-Gayatri patil
3- Brijesh Dwivedi

Problem Statement

- XEducationsells onlinecoursestoindustryprofessionals.
- XEducationgetsalotofleads,its leadconversionrateis verypoor.Forexample,if, say,theyacquire100 leads ina day,onlyabout 30ofthemareconverted.
- Tomakethisprocessmoreefficient,thecompanywishestoidentifythemost potentialleads,alsoknownas‘HotLeads’.
- If theysuccessfullyidentifythis setof leads,theleadconversionrateshouldgoupas thesalesteamwillnowbefocusingmoreoncommunicatingwiththepotentialleads ratherthanmakingcalls toeveryone.

Business Objective:

- X educationwantstoknow mostpromisingleads.
- Forthattheywanttobuilda Model whichidentifiesthe hotleads.
- Deploymentofthe model for the futureuse.

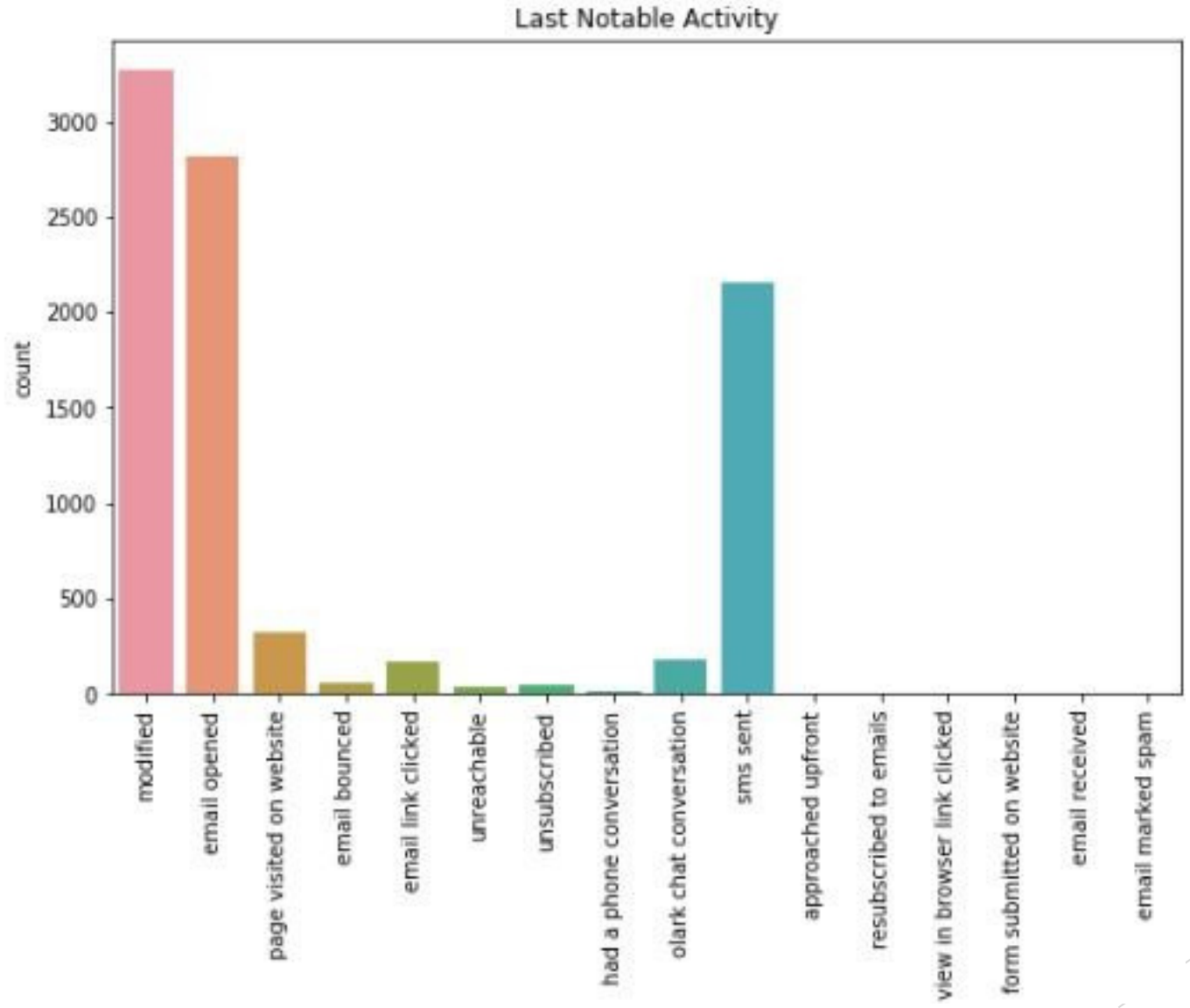
SolutionMethodology

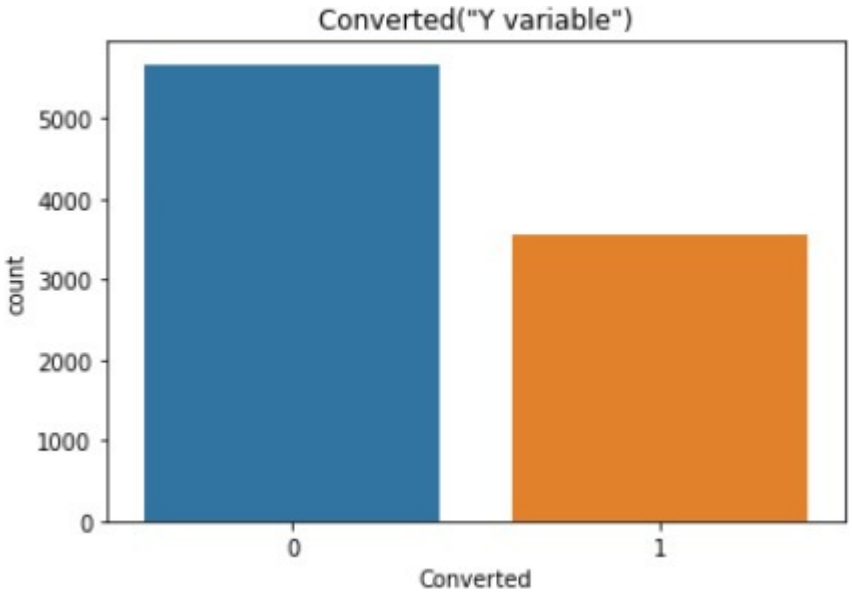
- Datacleaninganddatamanipulation.
  1.Checkandhandle duplicatedata.
  2.CheckandhandleNA valuesandmissingvalues.
  3.Dropcolumns,if itcontainslargeamountofmissing values andnotusefulfortheanalysis.
  4.Imputationof thevalues,ifnecessary.
  5.Checkand handleoutliers in data.
- EDA
  1.Univariate dataanalysis:valuecount,distributionofvariableetc.
  2.Bivariate dataanalysis: correlationcoefficientsandpatternbetweenthevariables etc.
- Feature Scaling&DummyVariablesand encodingofthe data.
- Classificationtechnique:logisticregressionusedforthemodelmakingandprediction.
- Validationofthemodel.
- Modelpresentation.
- Conclusionsandrecommendations.

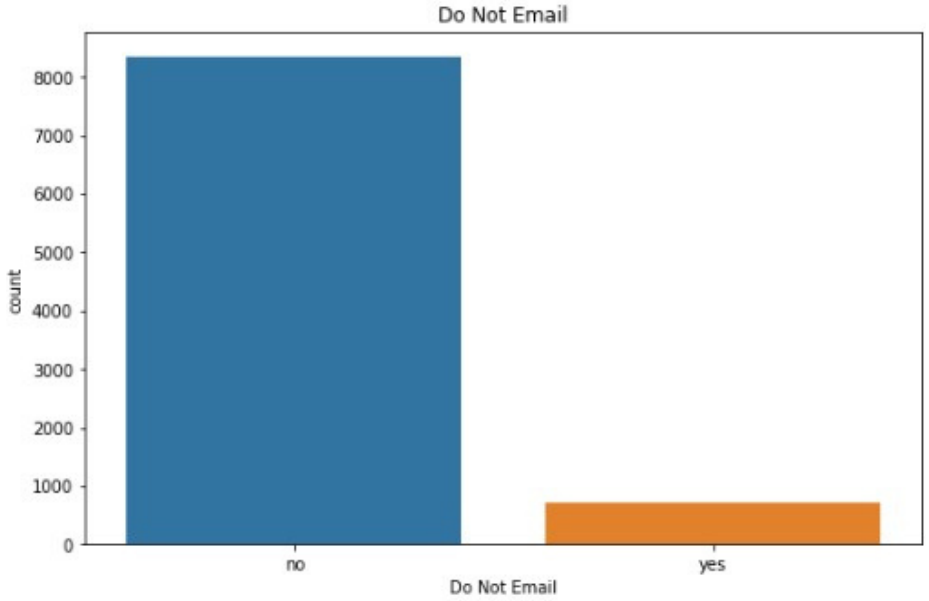
DataManipulation

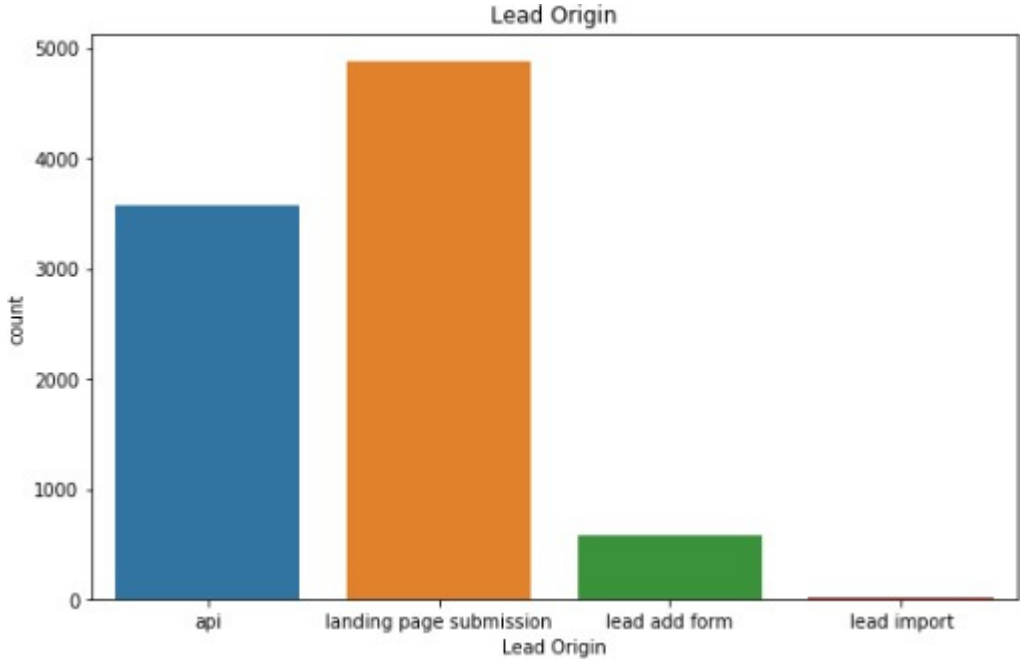
- TotalNumberofRows=37,TotalNumberofColumns=9240.
- Singlevaluefeatureslike“Magazine”,“ReceiveMoreUpdatesAboutOurCourses”, “UpdatemeonSupply”
- ChainContent”,“GetupdatesonDMContent”,“Iagreetopaytheamountthrough cheque”etc.havebeendropped.
- Removingthe“ProspectID”and “LeadNumber” whichisnotnecessaryfortheanalysis.
- Aftercheckingforthevaluecountsforsomeoftheobjecttypevariables,wefindsomeof thefeatureswhichhas noenoughvariance, whichwehavedropped,thefeaturesare: “DoNotCall”,“Whatmattersmosttoyouinchoosingcourse”,“Search”,“Newspaper Article”,“XEducation Forums”,“Newspaper”,“DigitalAdvertisement”etc.
- Droppingthecolumnshavingmorethan 35%asmissingvaluesuchas‘Howdid youhear aboutXEducation’ and‘Lead Profile’.

EDA

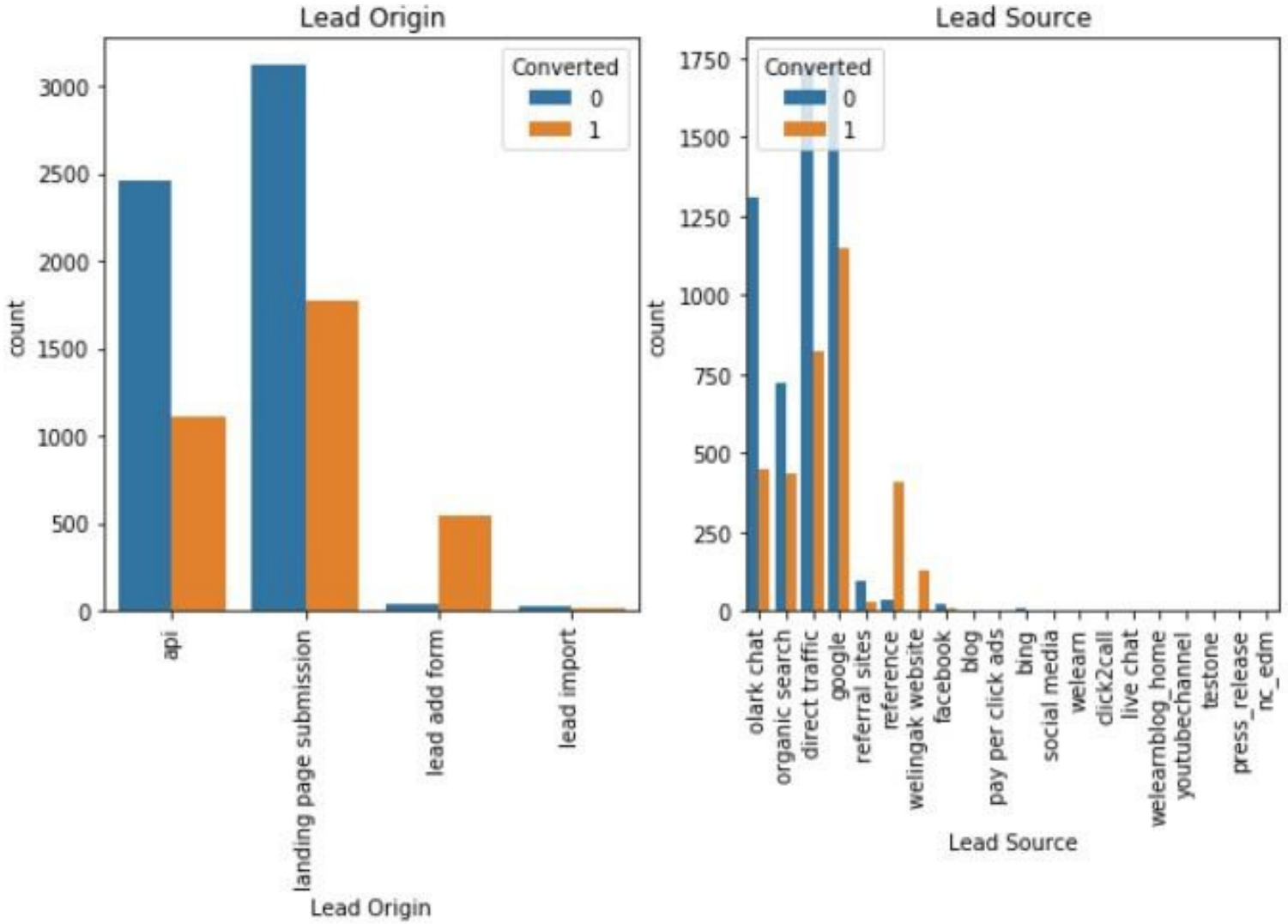

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam

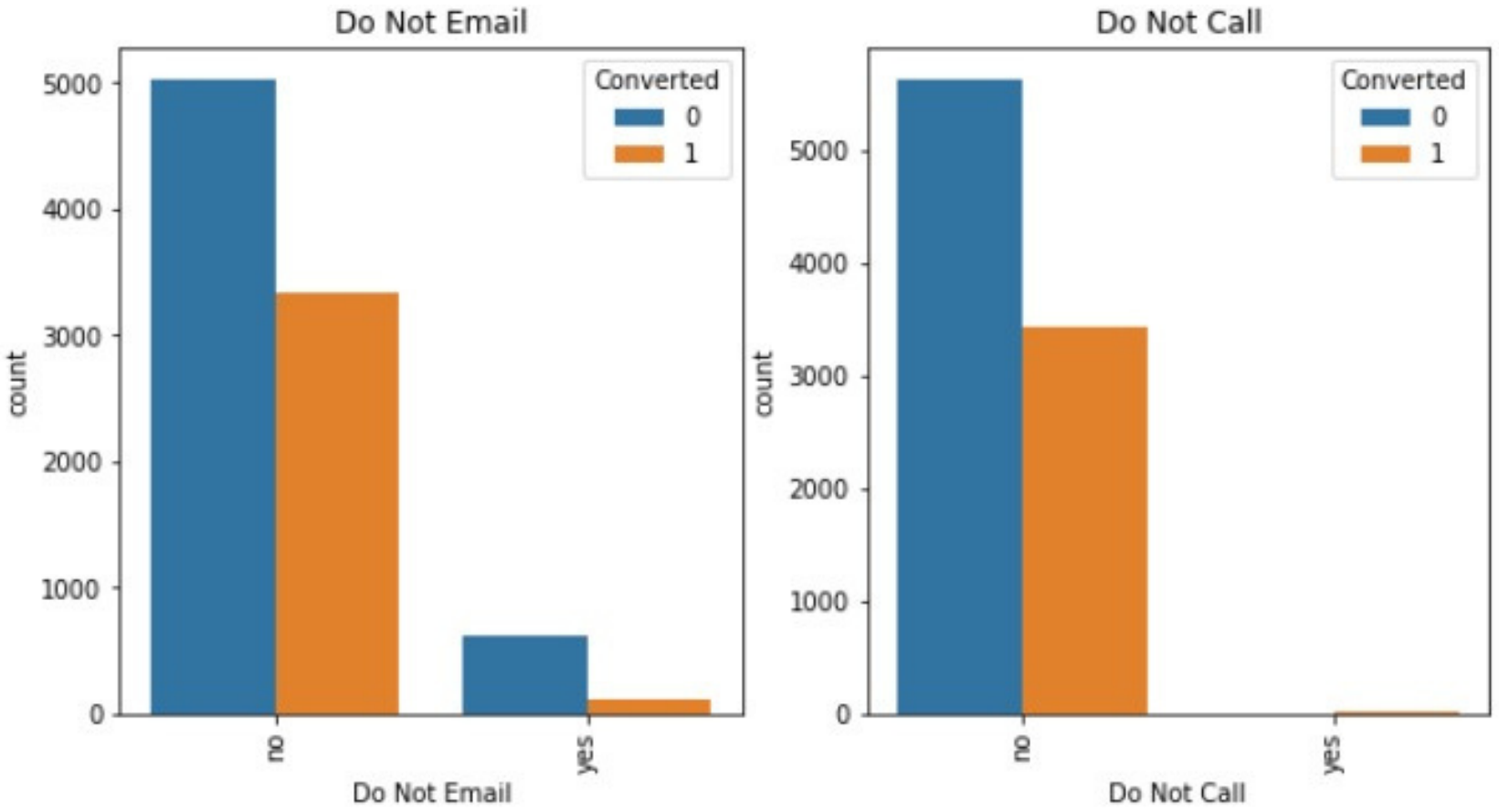

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

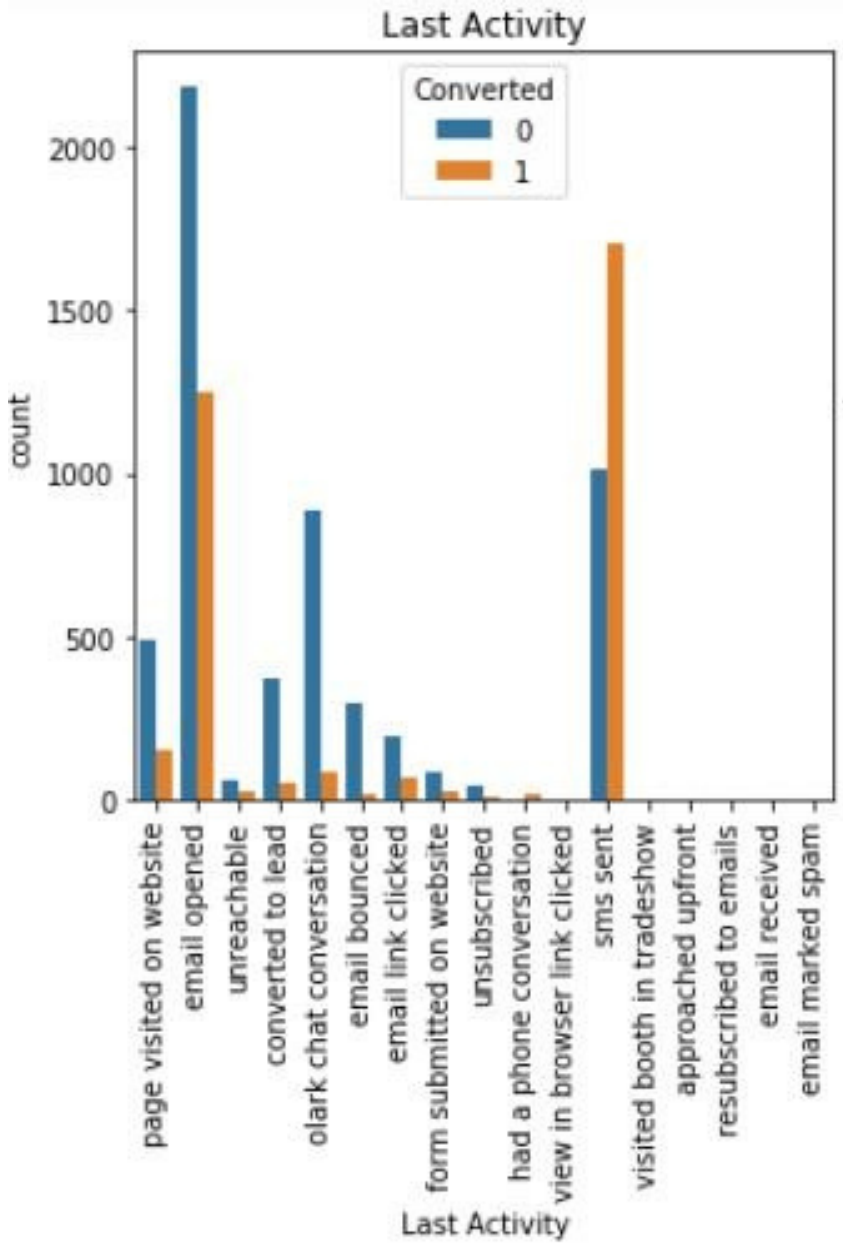

Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email

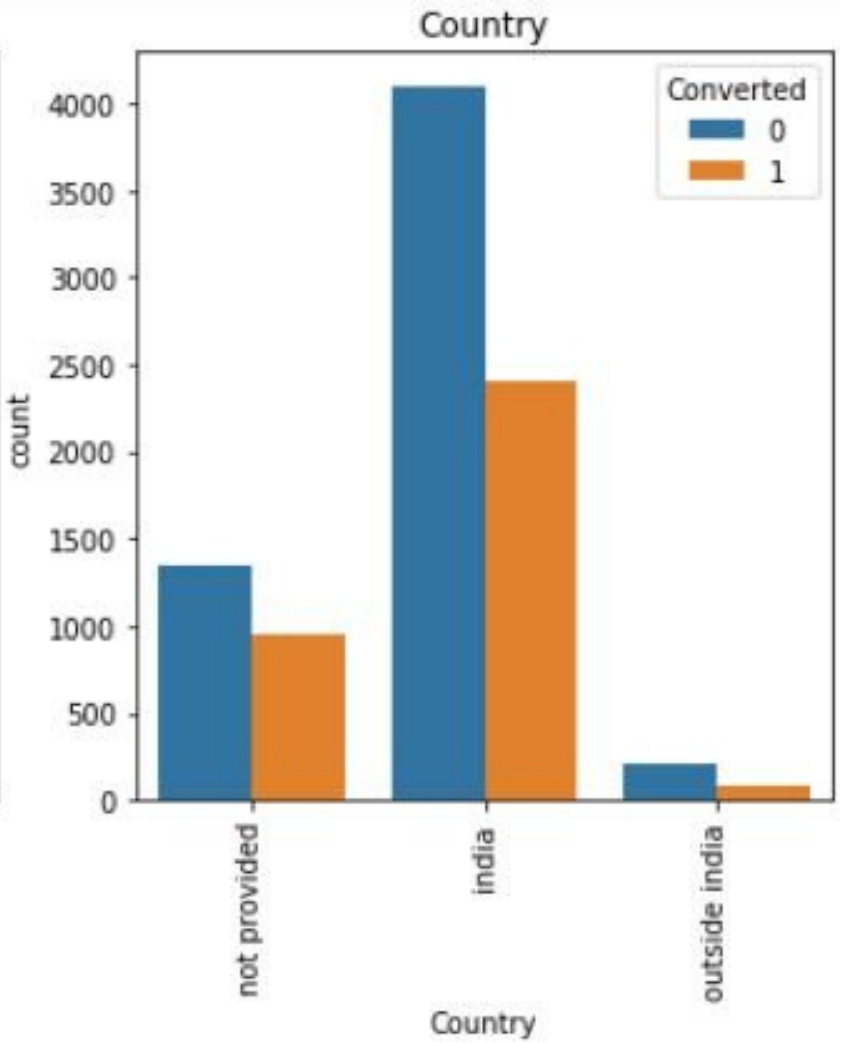

Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

CategoricalVariableRelation




Do Not Email
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Email
Do Not Call
Converted
0
1
5000
4000
3000
2000
1000
0
count
no
yes
Do Not Call


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity


Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

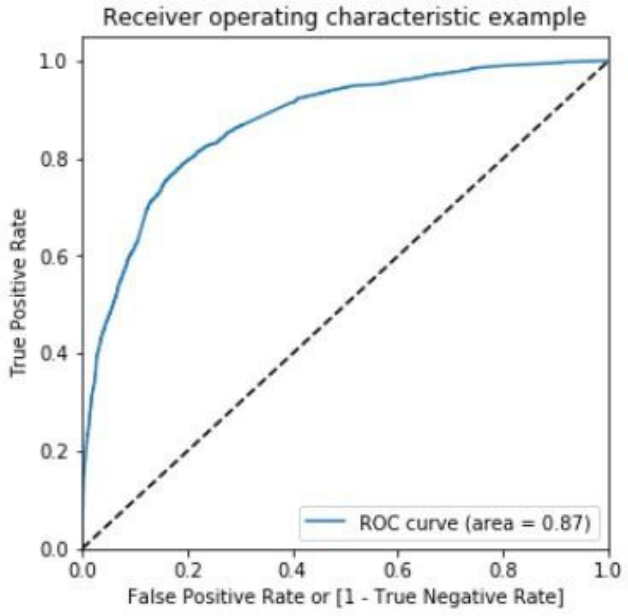
DataConversion

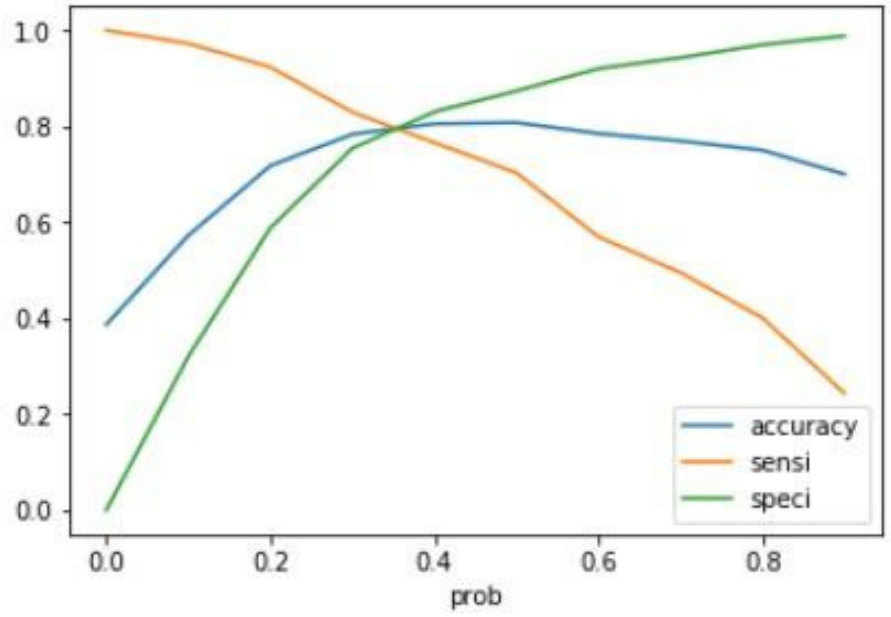
- NumericalVariablesareNormalised
- DummyVariablesarecreatedforobjecttypevariables
- TotalRows forAnalysis:8792
- TotalColumnsforAnalysis:43

ModelBuilding

- Splittingthe DataintoTrainingandTestingSets
- Thefirst basicstepforregressionisperforminga train-testsplit,wehavechosen70:30 ratio.
- UseRFEfor Feature Selection
  - Running RFEwith15variablesasoutput
  - Building Modelbyremovingthevariablewhosep-valueis greaterthan0.05 andvif valueisgreater than5
- Predictionsontestdataset
  - Overallaccuracy77%

ROCCurve





- **Finding Optimal Cut off Point**
- Optimalcutoff probabilityisthat
- probabilitywherewegetbalancedsensitivity andspecificity.
- From thesecondgraphit isvisiblethattheoptimalcutoff is at0.35.

Conclusion

Itwasfoundthatthevariablesthatmatteredthemostin thepotentialbuyersare(In descendingorder):

- The totaltimespendonthe Website.
- Totalnumberofvisits.
- Whenthelead sourcewas:
  a.Google
  b.Directtraffic
  c.Organicsearch
  d.Welingakwebsite
- Whenthelastactivitywas:
  a.SMS
  b.Olarkchat conversation
- When theleadorigin isLeadadd format.
- Whentheircurrentoccupationis asa workingprofessional.

Keepingthese inmindtheXEducationcanflourishastheyhaveaveryhigh chancetogetalmostallthepotentialbuyerstochangetheirmindandbuytheir courses.