LEAD SCORING CASE STUDY

Group details:

1. Sai Tharun
2. Mujtaba
3. Yashasvi

PROBLEM STATEMENT

1. X Education offers online courses to industry professionals.

2. Despite getting many leads, the conversion rate is low. For example, out of 100 daily leads, only about 30 convert.

3. To improve efficiency, X Education wants to identify 'Hot Leads'—the most promising leads.

4. By focusing on these potential leads, the sales team can increase the conversion rate by targeting their efforts more effectively.

5. Business Objective:
   - X Education wants to identify the most promising leads.
   - They aim to build a model to identify these hot leads.
   - The model will be deployed for future use.

SOLUTION METHODOLOGY

- Data Cleaning and Manipulation:
    1. Handle Duplicates: Identify and remove any duplicate entries in the data.
    2. Address NA and Missing Values: Check for NA and missing values and handle them appropriately.
    3. Drop Unnecessary Columns: Remove columns with a large amount of missing data that are not useful for analysis.
    4. Impute Values: Perform imputation for missing values where necessary.
    5. Outlier Management: Identify and manage outliers in the data.

- Exploratory Data Analysis (EDA):
    1. Univariate Analysis: Conduct univariate analysis to check value counts and distribution of variables.
    2. Bivariate Analysis: Perform bivariate analysis to examine correlation coefficients and patterns between variables.

- Feature Scaling and Encoding: Scale numerical features and create dummy variables for categorical data as needed.

- Model Building: Use logistic regression for classification and prediction.

- Model Validation: Validate the model to ensure accuracy and reliability.

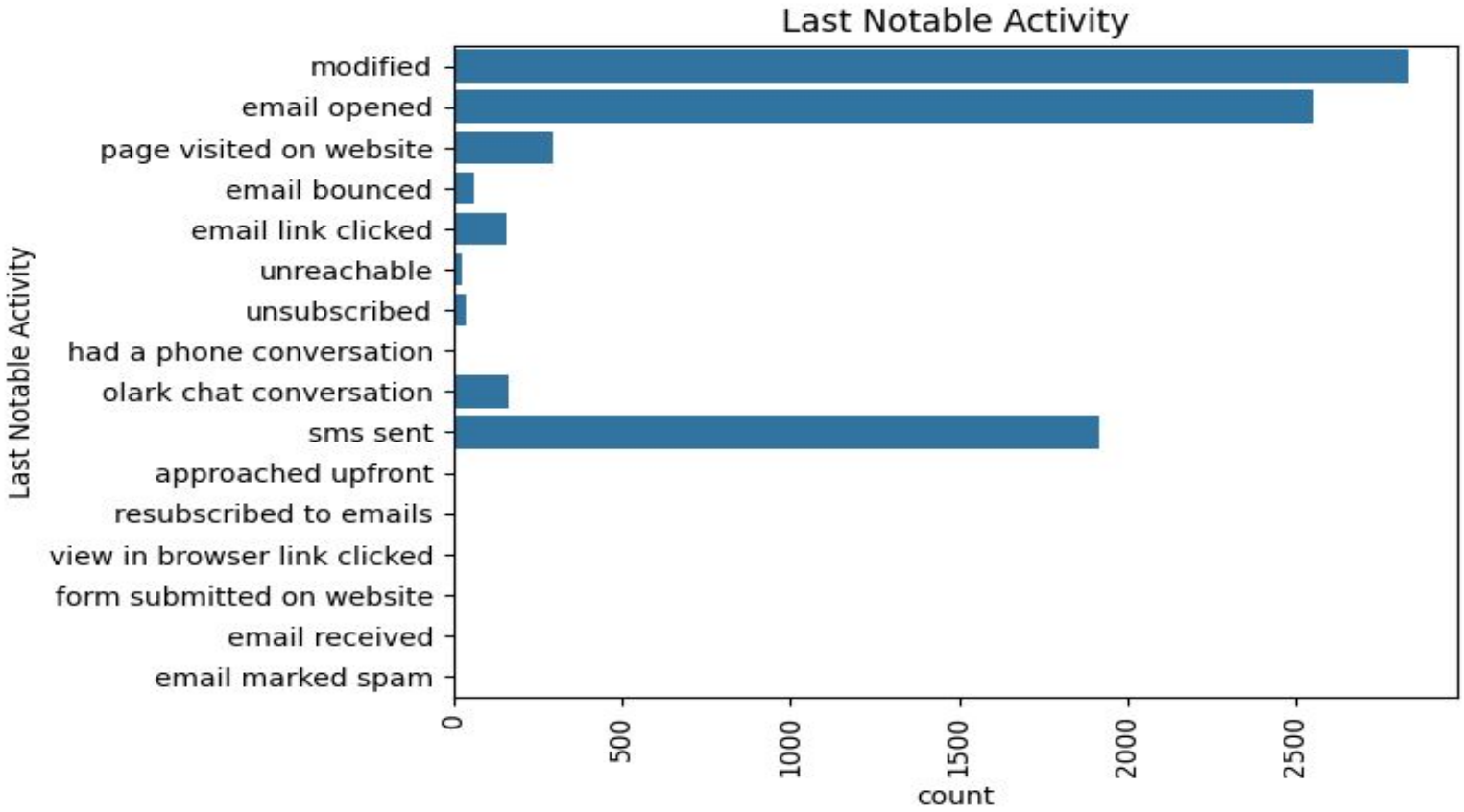
- Model Presentation: Present the model findings clearly.

- Conclusions and Recommendations: Provide conclusions based on the analysis and recommend actionable steps.

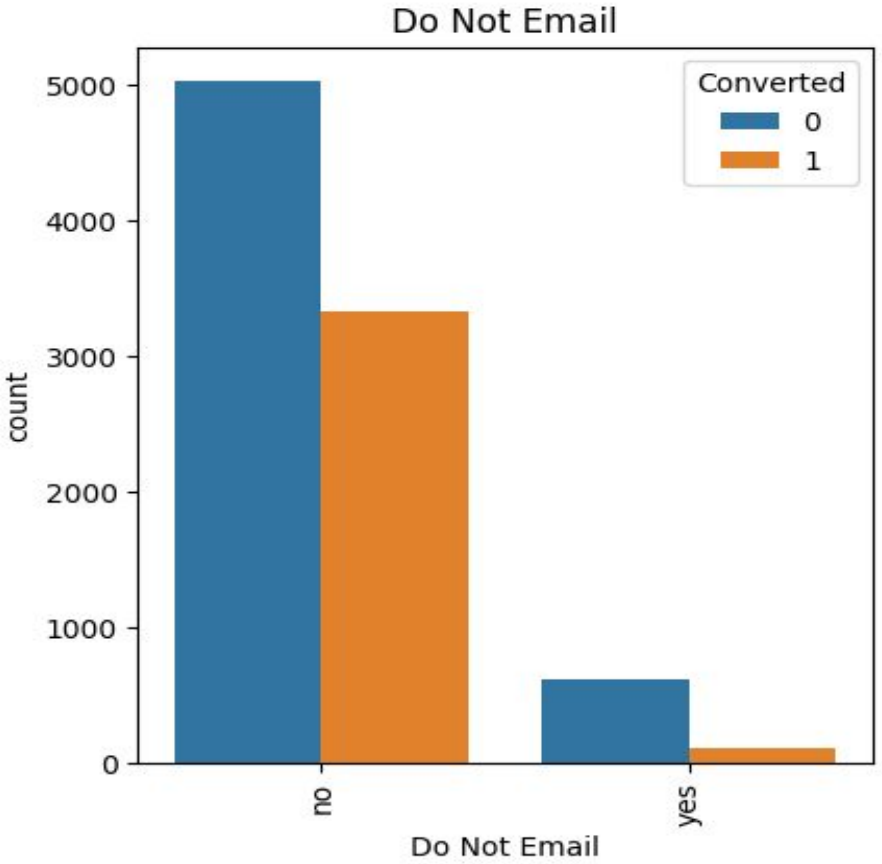
DATA MANIPULATION

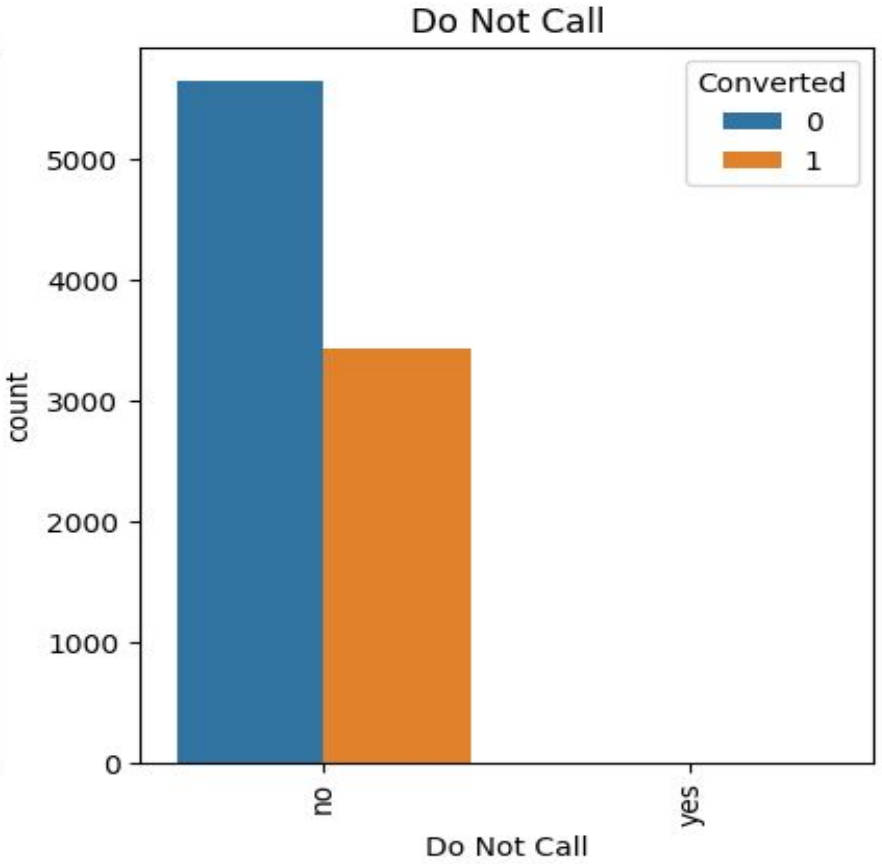
- Dataset Overview: The dataset consists of 37 rows and 9,240 columns.
- Drop Single-Value Features: Features with only a single value, such as “Magazine,” “Receive More Updates About Our Courses,” and “Update me on Supply,” have been removed.
- Irrelevant Features Removed: Columns like “Chain Content,” “Get updates on DM Content,” and “I agree to pay the amount through cheque” have been excluded.
- Unnecessary Identifiers Removed: "Prospect ID" and "Lead Number" were removed as they are not needed for analysis.
- Low-Variance Features Dropped: Features with insufficient variance, such as “Do Not Call,” “What matters most to you in choosing course,” “Search,” “Newspaper Article,” “X Education Forums,” “Newspaper,” and “Digital Advertisement,” were removed after analyzing value counts for object type variables.
- High Missing Value Columns Dropped: Columns with more than 35% missing values, like ‘How did you hear about X Education’ and ‘Lead Profile,’ were also dropped.

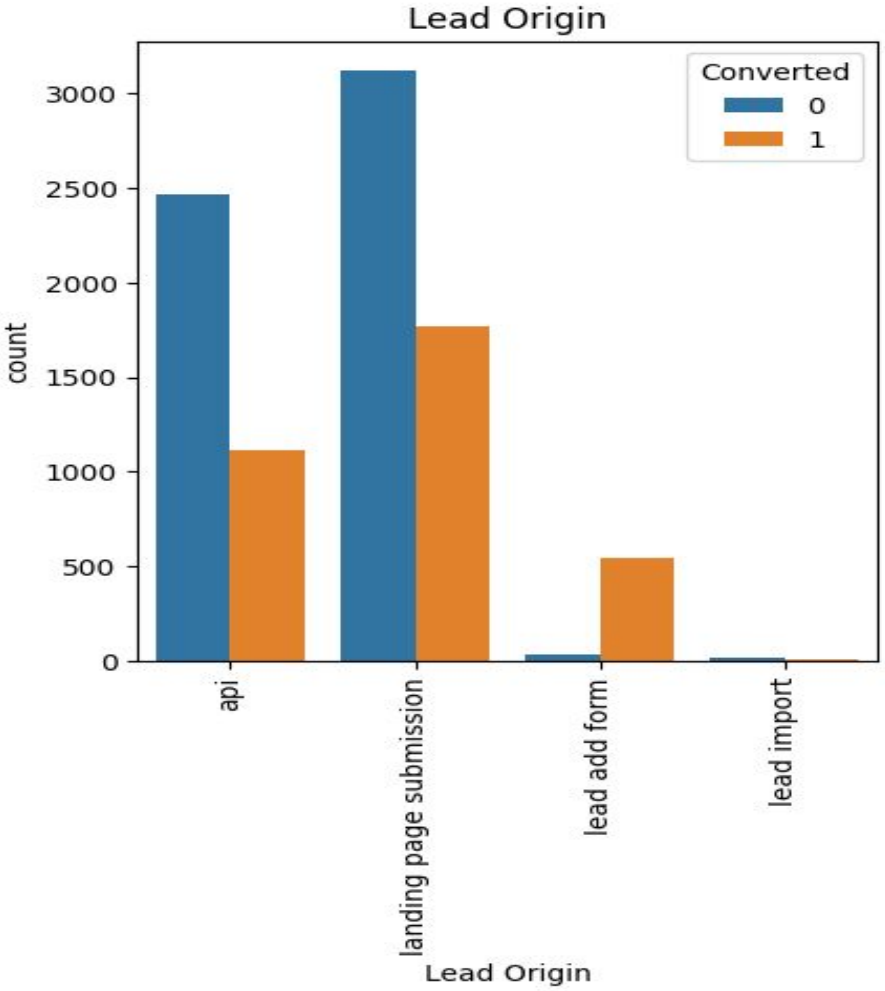
Exploratory Data Analysis

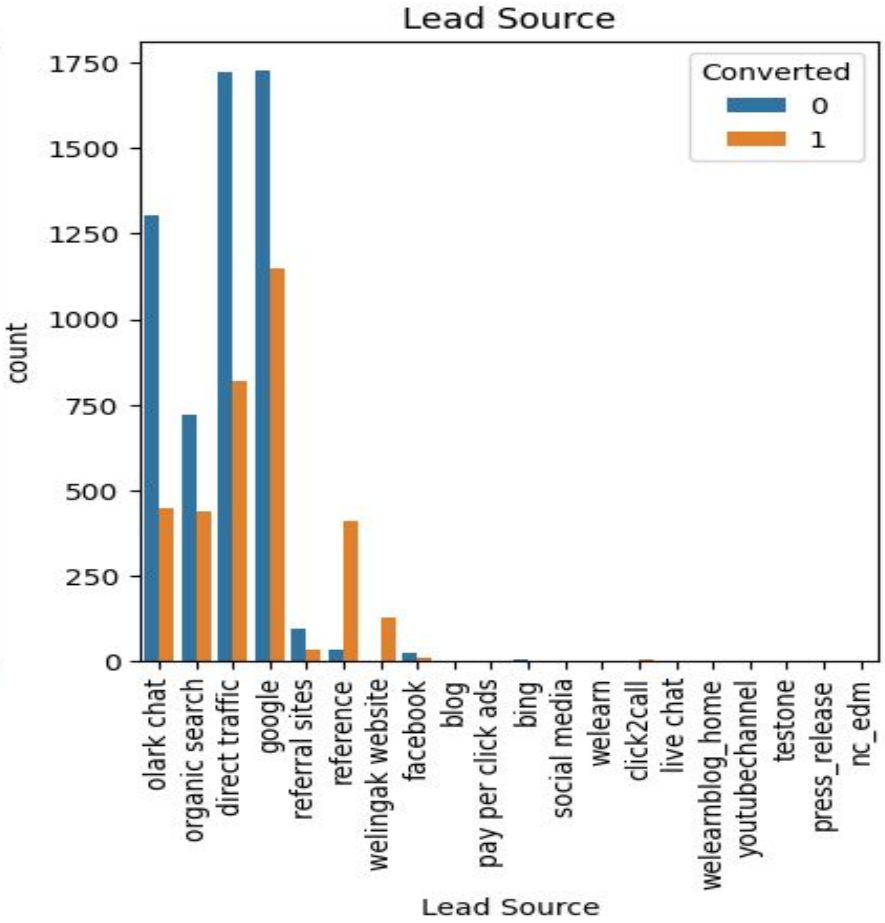


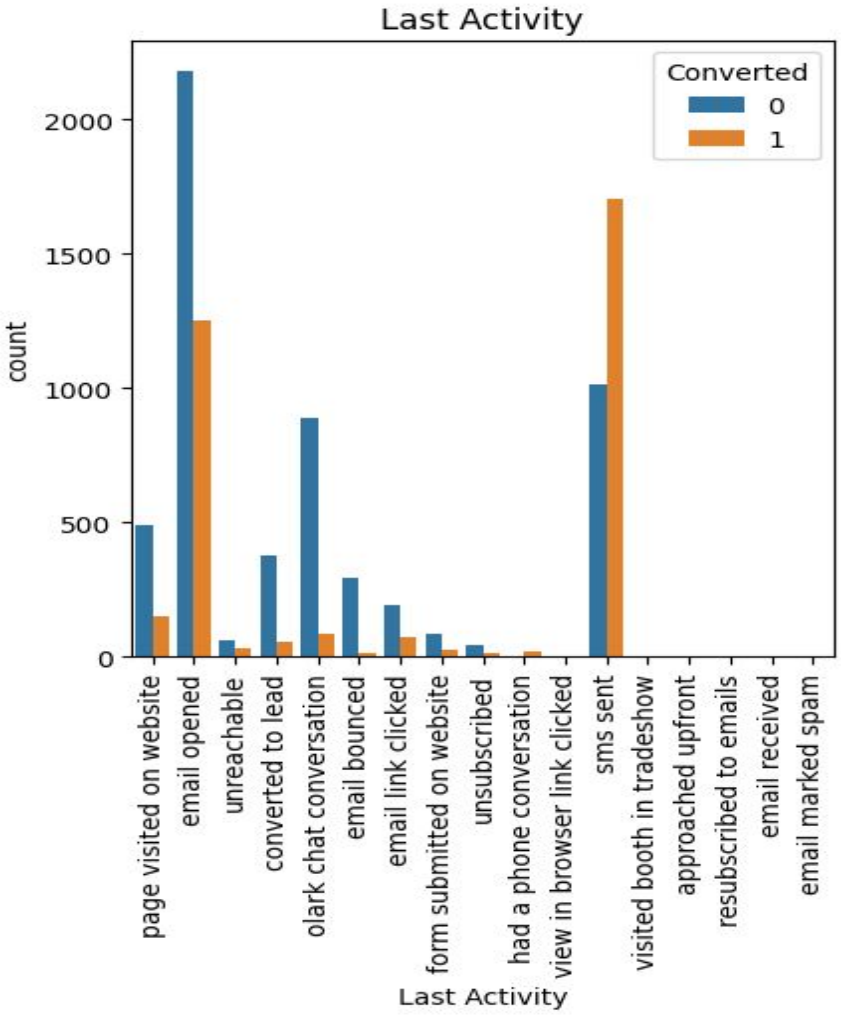
CATEGORICAL VARIABLE RELATION

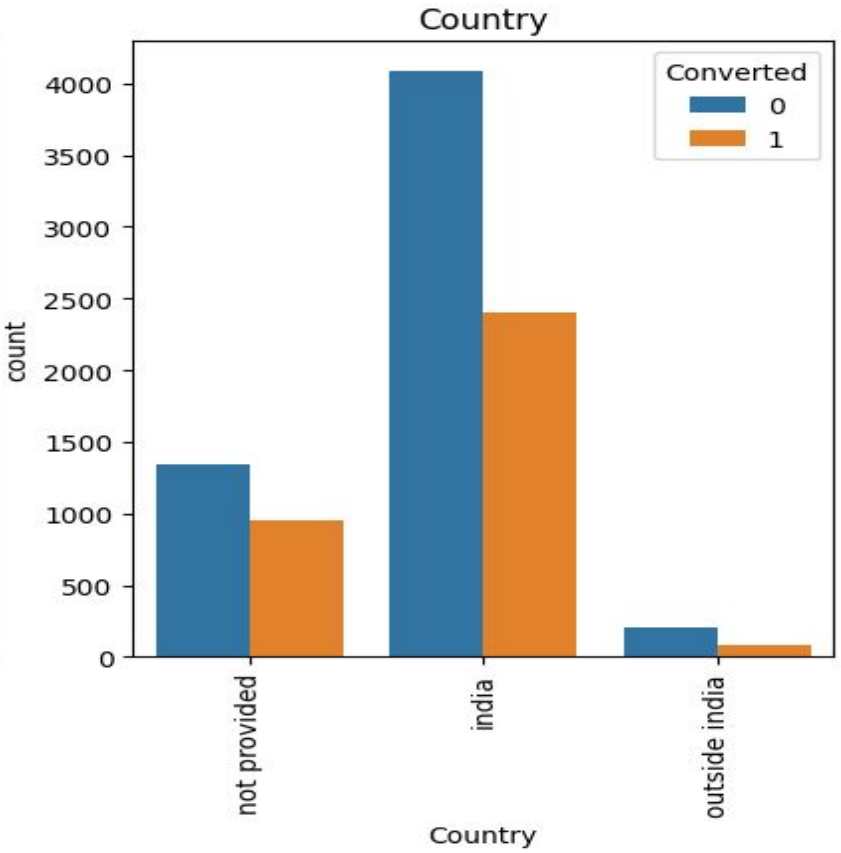





Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
api
landing page submission
lead add form
lead import
Lead Origin


Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source


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

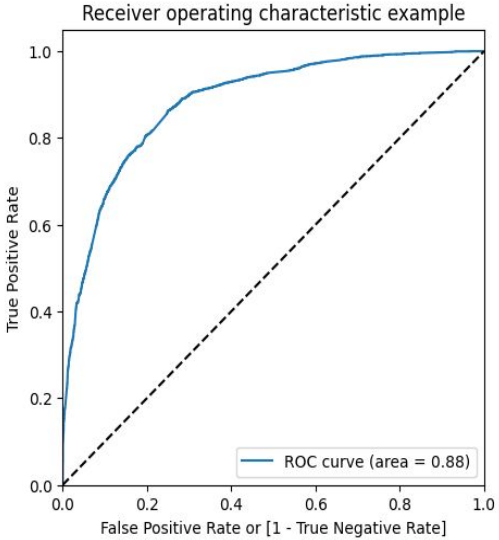
Data Conversion:

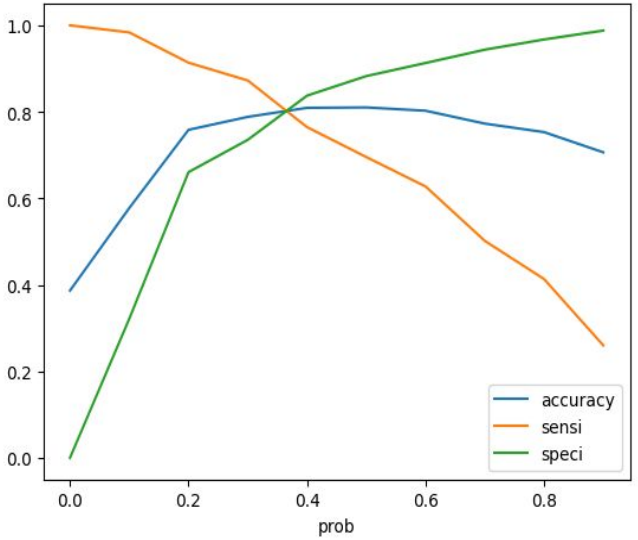
- Normalization of Numerical Variables: Numerical variables have been normalized.
- Creation of Dummy Variables: Dummy variables were created for categorical (object type) variables.
- Dataset Size for Analysis: The dataset now consists of 8,792 rows and 43 columns for analysis.

Model Building

- Train-Test Split: The data is split into training and testing sets in a 70:30 ratio.
- Feature Selection with RFE: RFE is run to select 15 important variables.
- Model Refinement: Variables with a p-value greater than 0.05 and VIF greater than 5 are removed.
- Predictions: The model makes predictions on the test data.
- Accuracy: The overall accuracy of the model is 81%.

ROC CURVE





- Optimal Cutoff Point: Identify the probability that balances sensitivity and specificity.
- Cutoff Value: The optimal cutoff probability is determined to be 0.35, as shown in the second graph.

Conclusion

It turns out that the factors that mattered the most to potential customers were (in descending order):

1) The total time spend on the Website.
2) Total number of visits.
3) When the lead source was:
    - Google
    - Direct traffic
    - Organic search
    - Welingkar website
4) When the last activity was:
    - SMS
    - Olark chat conversation
5) When the lead origin is Lead add format.
6) When their current occupation is as a working professional.

If X Education keeps these in mind, they stand a very good chance of convincing nearly every prospective customer to change their mind and enroll in their courses.

Thank you