Lead scoring

Himanshu saluja

Problem statement

X Education, an online course provider, faces low lead conversion despite a high volume of leads. With a current conversion rate of 30%, far from the desired 80%, they aim to identify 'Hot Leads' with potential for higher conversion. To achieve this, they plan to develop a lead scoring model. This model will evaluate factors like website interaction, form submissions, referrals, and professional status to prioritize leads likely to convert. By focusing efforts on these promising leads, they aim to increase the overall conversion rate and align with the CEO's target. Through targeted communication and nurturing, X Education aims to improve lead quality and achieve sustainable growth.

Goal

The main goals of this case study are:

1. Develop a logistic regression model to assign lead scores ranging from 0 to 100, indicating the likelihood of conversion. Higher scores represent hotter leads with a higher chance of conversion, while lower scores indicate colder leads with less chance of conversion.
2. Ensure that the model can adapt to future changes in company requirements by addressing additional problems provided by the company. These issues should be incorporated into the logistic regression model's framework to maintain its effectiveness over time. This information should also be included in the final presentation for making recommendations.

Steps

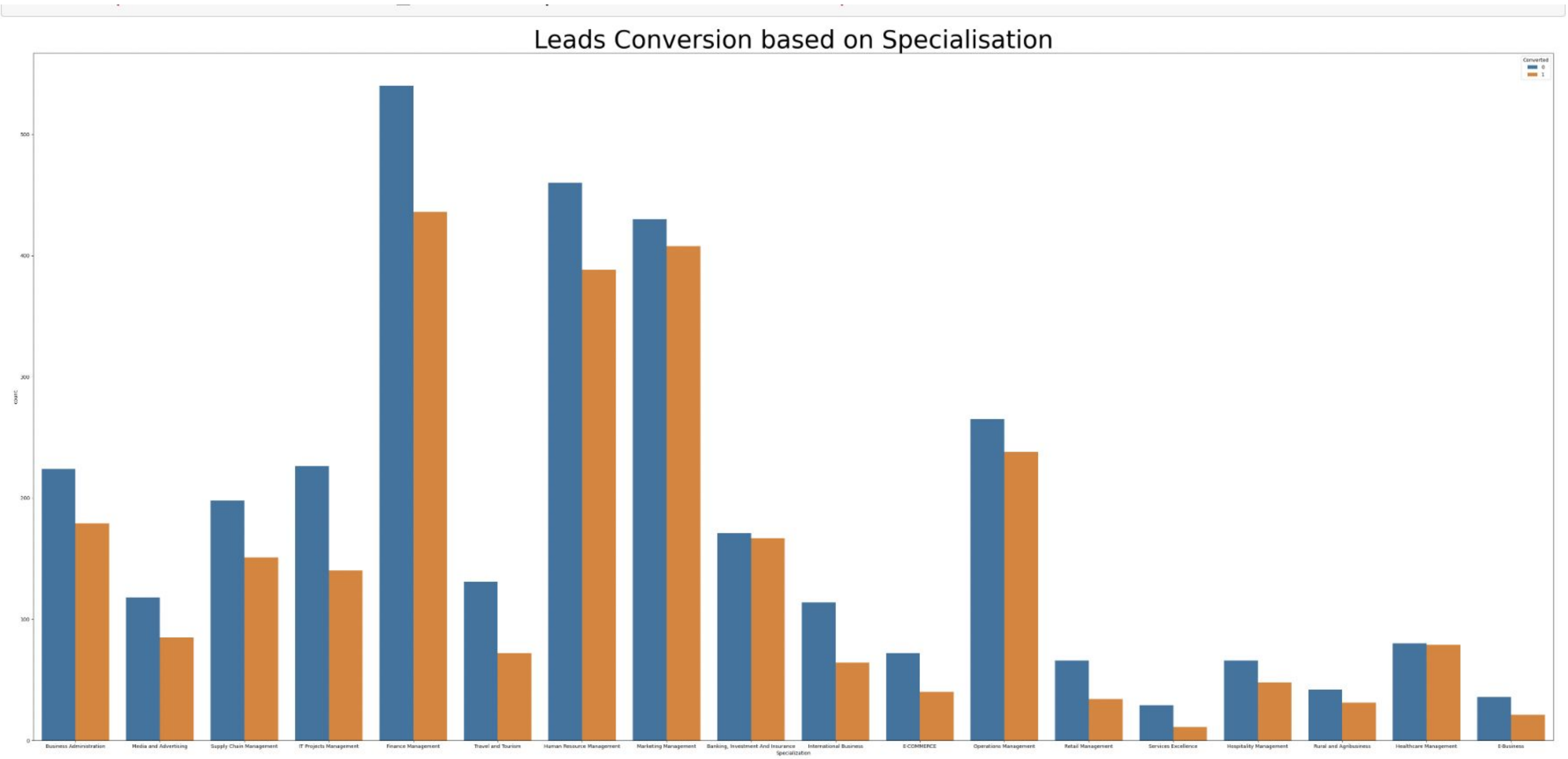
1. Data cleaning and manipulation
2. EDA
    a. Univariate data analysis
    b. Bivariate data analysis
    c. Outlier detection and fixing
    d. Removing null or empty values
3. Feature scaling and dummy variable
4. Logistic regression
5. ROC curve
6. Perdiction
7. Conclusions

Data manipulations

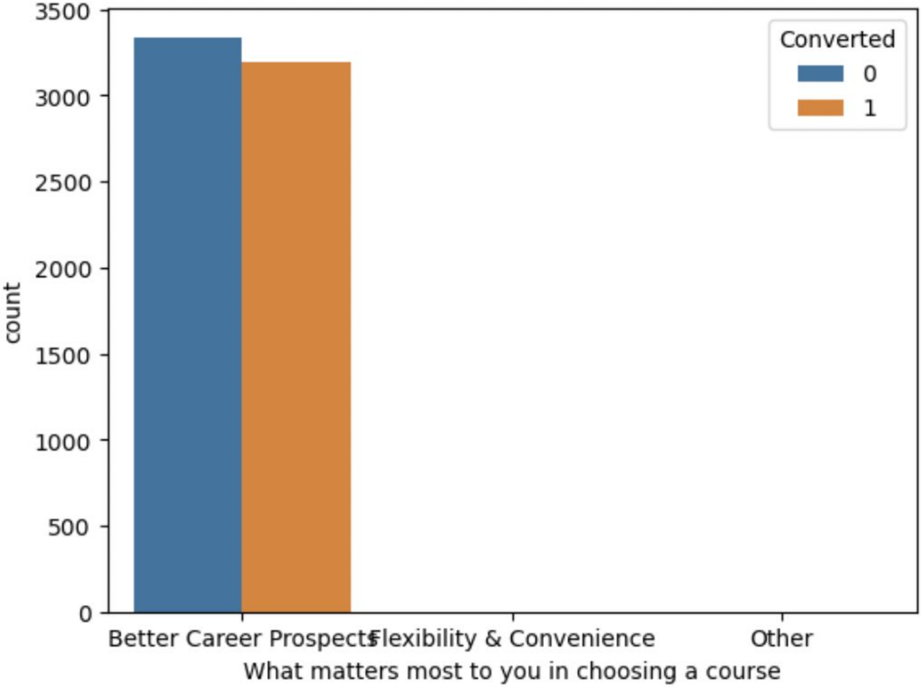
1. Location data city, country is more skewed towards one value, and the majority values are null too so droping these columns should be fine
2. Not so useful or manually created column that are filled by lead agent later can also be removed - ['Lead Source','Last Notable Activity','Tags','Last Activity']
3. Columns with greater than 40% null values can also be removed - ['Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score']
4. 100% of value lead to only 2 data points , we can remove cols - ‘What matters most to you in choosing a course’
5. 90% of data is either for NAN or unemployed - removing - What is your current occupation
6. ['Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'] - these columns can be dropped since all of them have 1 same value and no null values
7. ['Do Not Call','Search', 'Newspaper Article','X Education Forums', 'Newspaper','Digital,Advertisement','Through Recommendations'] these col are biased towards 1 value by 90% dropping them

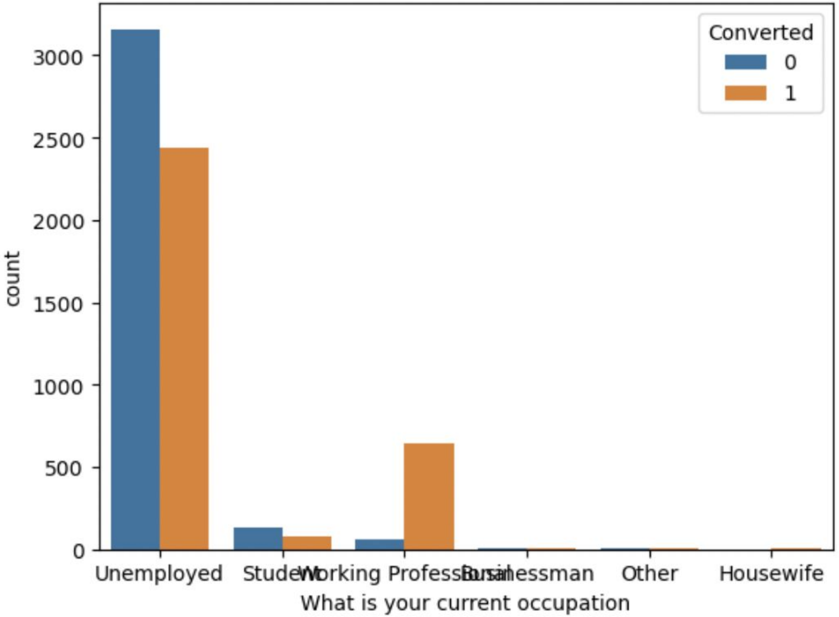
EDA

Categorical variable

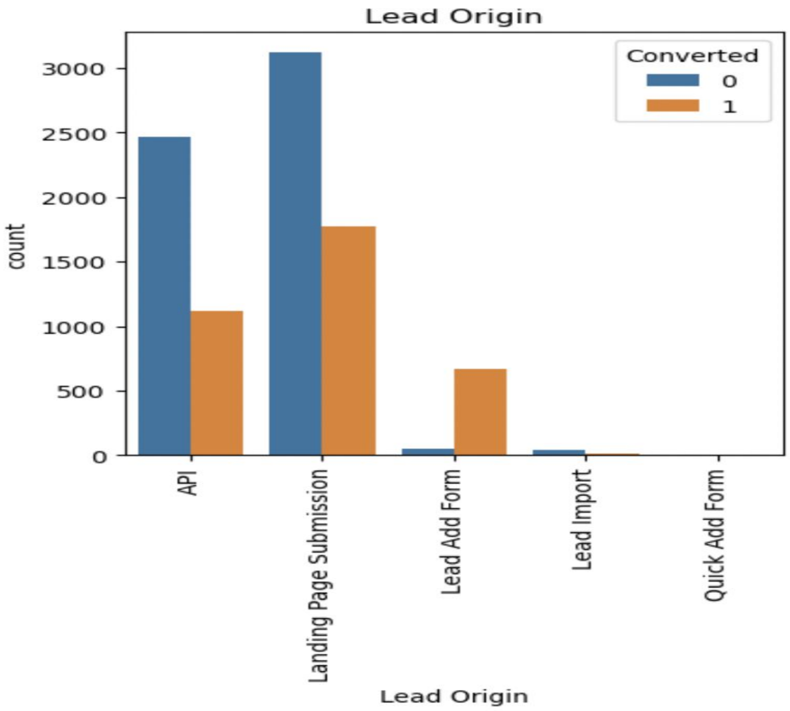


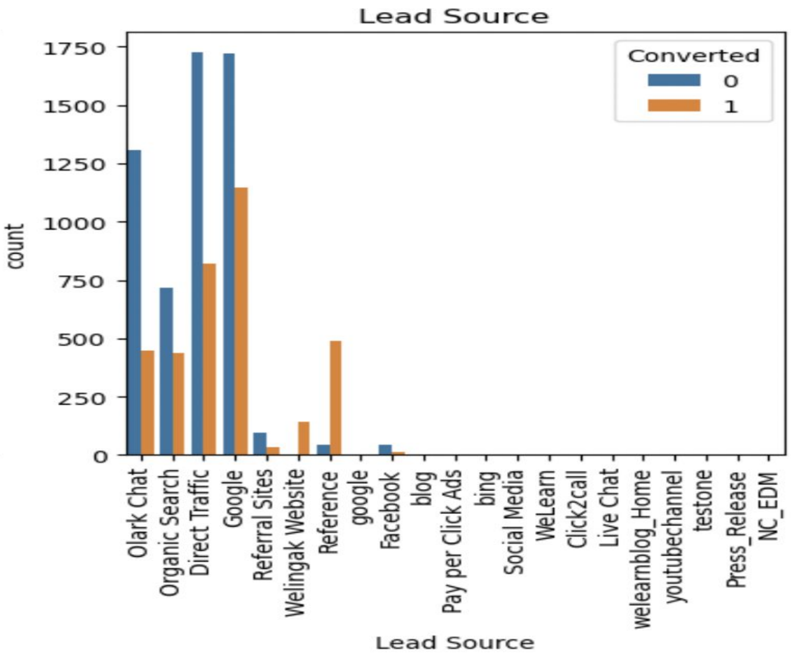
Categorical variable



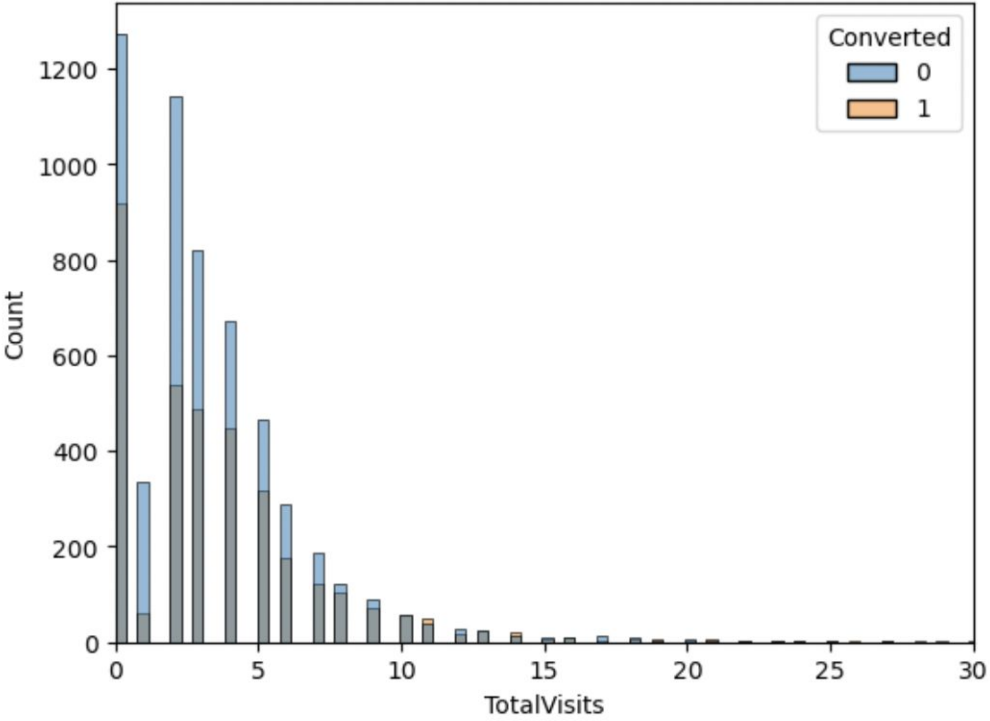


Categorical variable

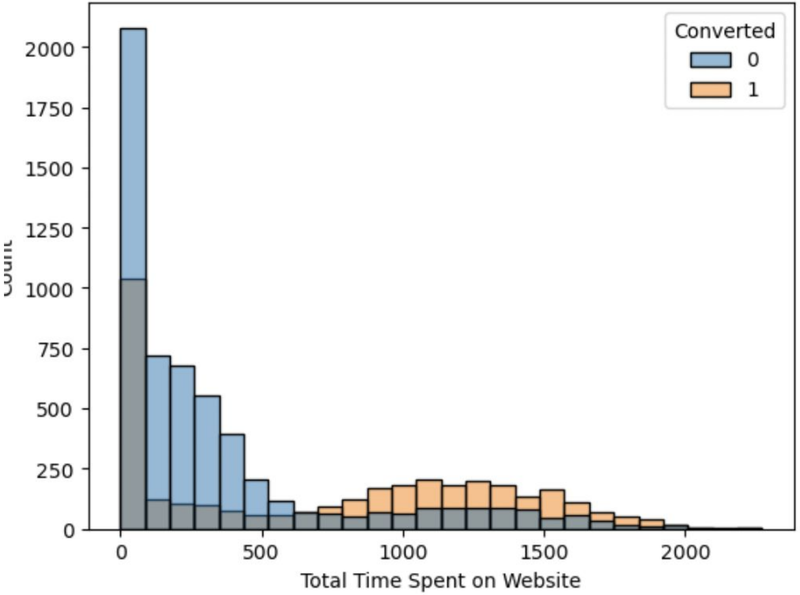


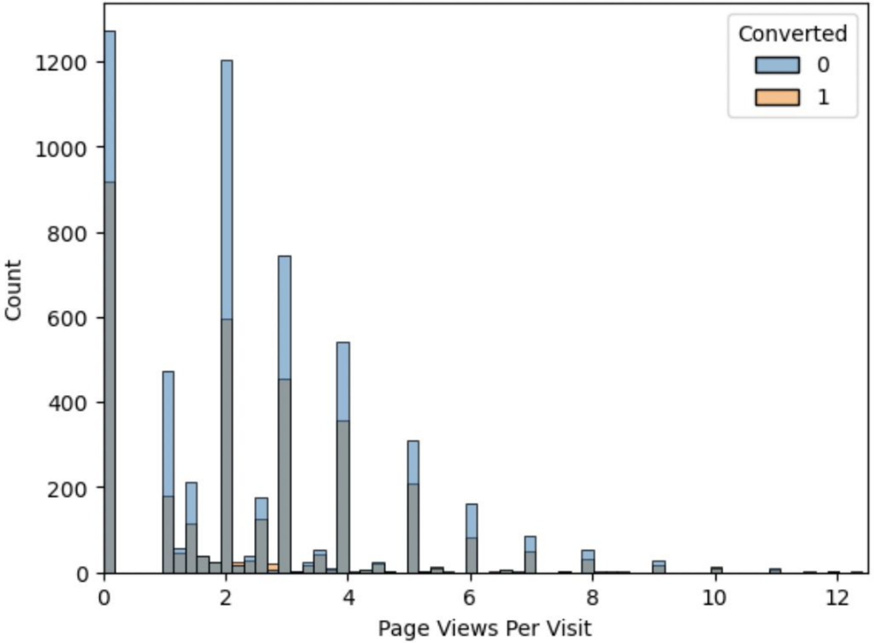


Numerical variable analysis

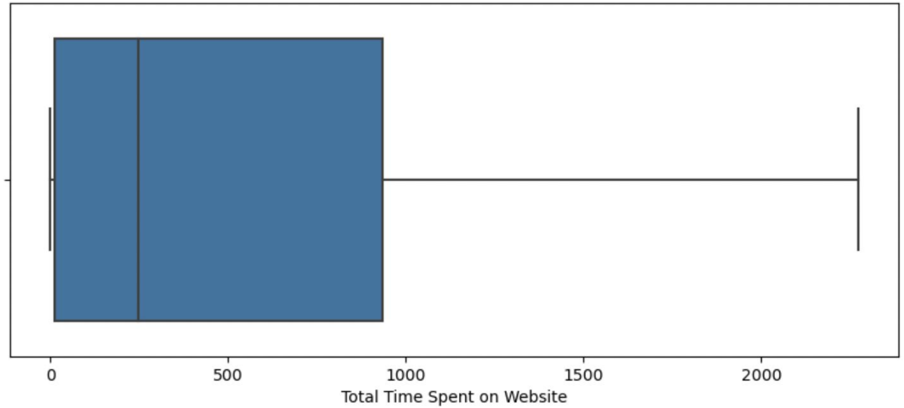


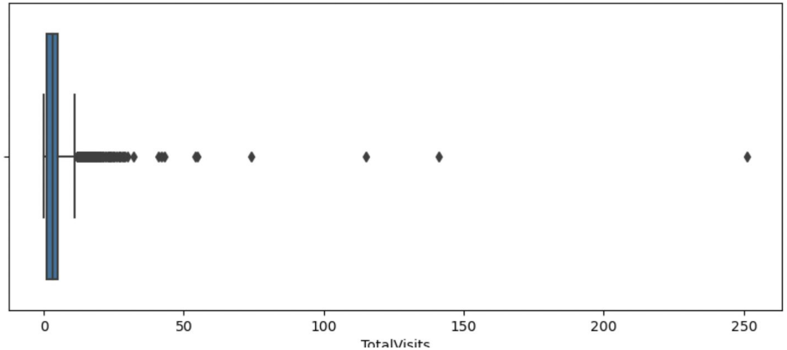
Numerical variable analysis

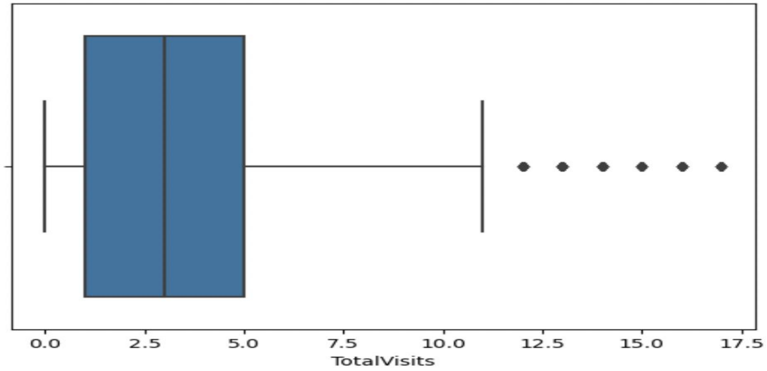




Outliner checking







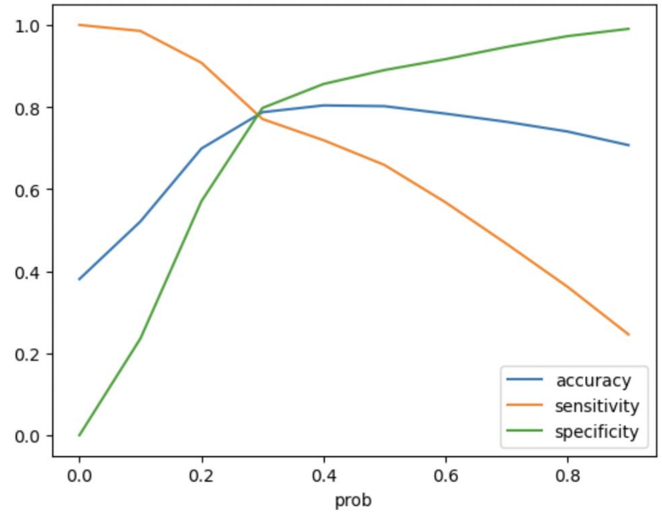
Model building

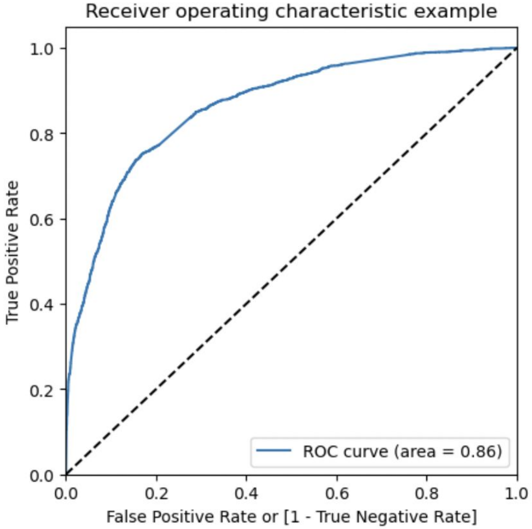
- Split data in train vs test set
- Use RFE for feature selection
- Build model by choosing 15 variable
- Remove variable with p value greater than 0.05
- Check for VIF to be less than 10
- Predict on test set
- Calculate accuracy

Numbers

- Training set
  - Accuracy - 80.19 %
  - Sensitivity is 65.87 %
  - Specificity is 89.0 %
- Test set
  - Accuracy - 80.11%
  - Sensitivity is 74.93 %
  - Specificity is 83.31 %

ROC curve


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
prob
accuracy
sensitivity
specificity


Receiver operating characteristic example
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.86)

Conclusion

1. Lead Origin in Lead Add Form convert mote
2. Working profession tend to convert more
3. Lead source - with following convert more
   a. Lead Source_Welingak Website
   b. Lead Source_Olark Chat
   c. Lead Source_Social Media
4. Total Time Spent on Website - this also increase conversion chances