# Building a Predictive Model for LEAD SCORE Optimization

1

Submitted By,

RITIKA, RAVI AND ROHINI

# CONTENT

# Problem Statement



An education company named X Education sells online coursesto industry professionals. On any given day, many professionals who are interested in thecourses land on their website and browse for courses.

The overall objective is to increase the enrollment of student by finding the important features

# Data Import



- Importing dataset
- Check the head of our master dataset
- Check the dimensions of the dataframe
- look at the statistical aspects of the dataframe
- Chek the type of each column

# Exploratory Data Analysis



- **Data Preparation**
- Check the percentage of null values in each column and drop the columns with a unique number, such as Prospect ID and Lead Number.
- Convert NaN values to 'Select'
- drop those columns with more than 45% missing values.
- Replace all null values with "Not Provided".

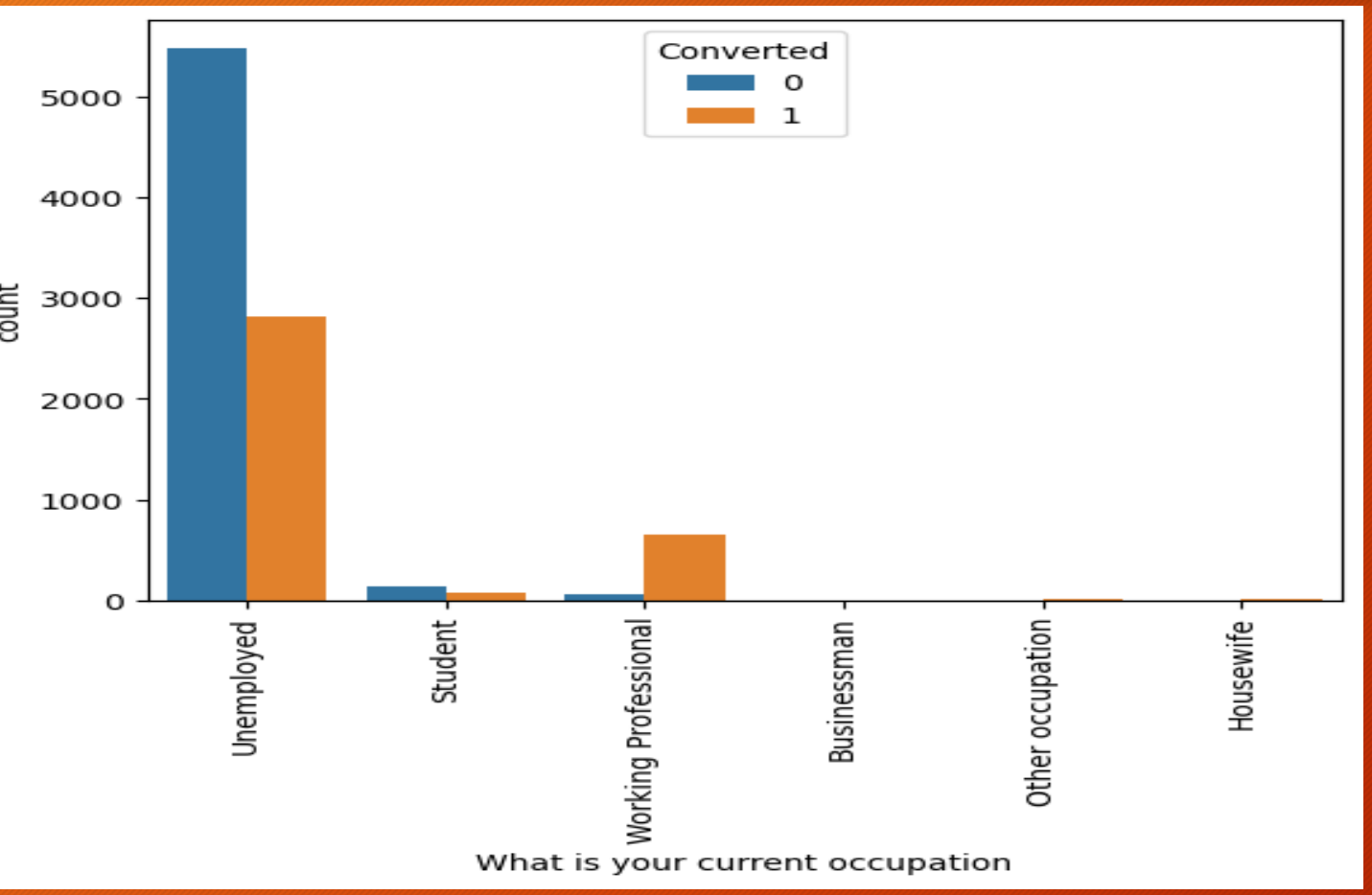
# Categorical Variable Analysis



- This text discusses the analysis of categorical variables for a dataset.
- Not provided values in the country and Nan values are replaced with India and Mumbai respectively.
- Small values from the Lead source column are replaced with "Others_Lead_Source".
- The "Do Not Email" column is dropped as 94% of its values give a "No" response, and
- columns with one value having more weight than 95% are also dropped.
- The Nan values of the current occupation column are imputed with the mode "Unemployed".
- Finally, the column "What matters most to you in choosing a course" is dropped as it does not contribute much.

# Categorical Variable Analysis



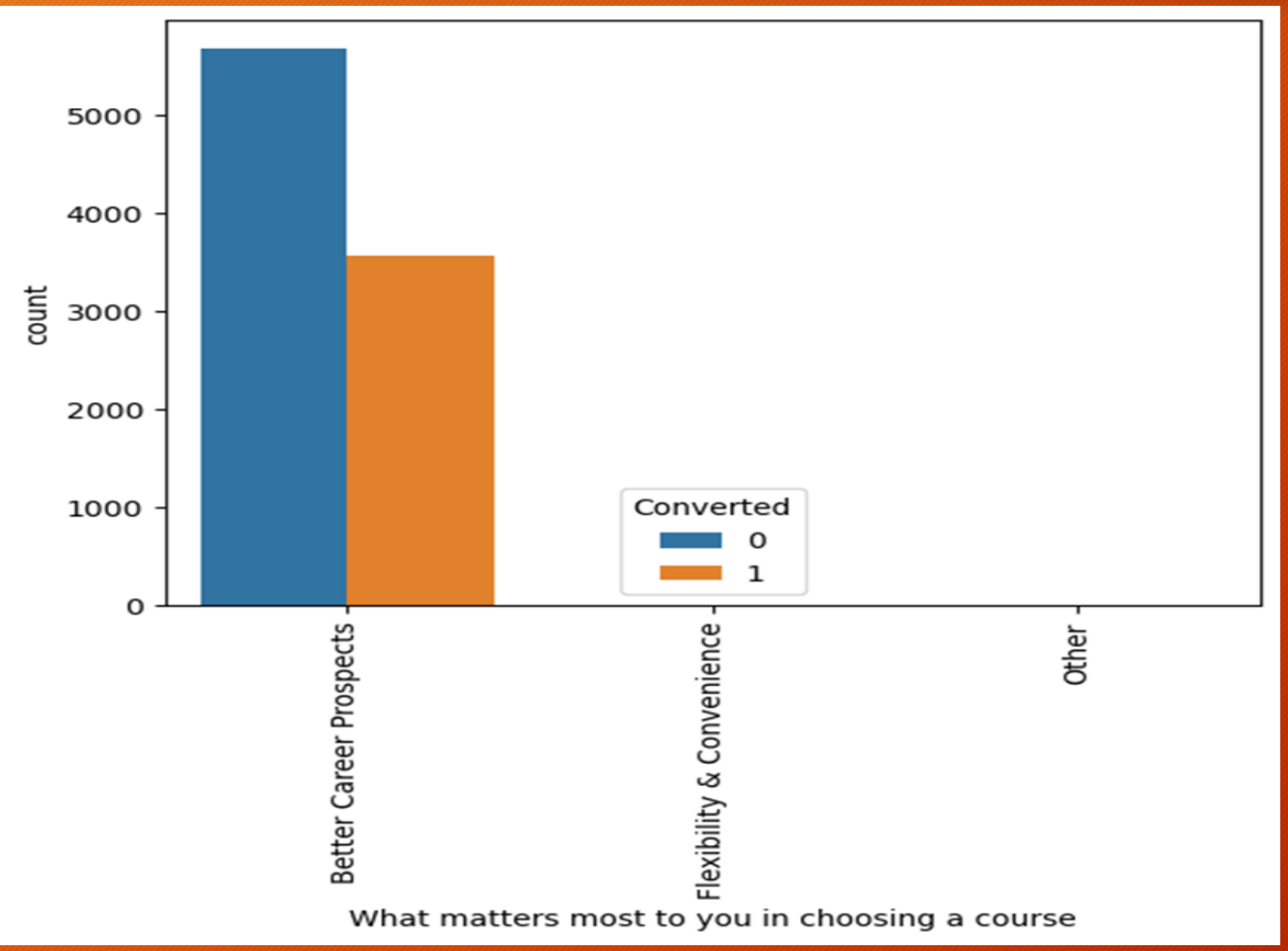



1) Working Professionals going for the course have high chances of joining it.

2) Unemployed leads are the most in terms of Absolute numbers.

# Categorical Variable Analysis

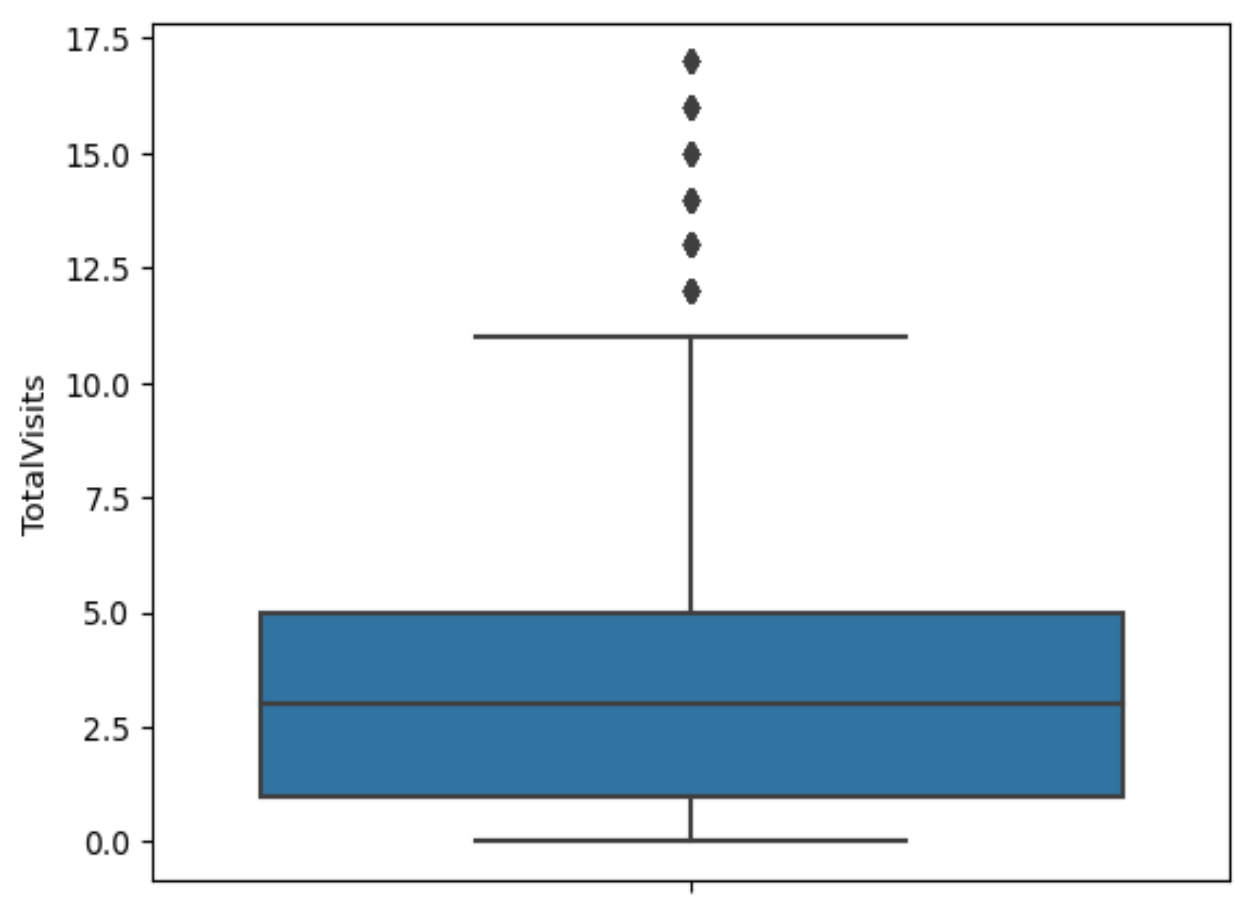





**The column 'What matters most to you in choosing a course' can be dropped as it doesn't contribute much**

# Numerical Variable Analysis





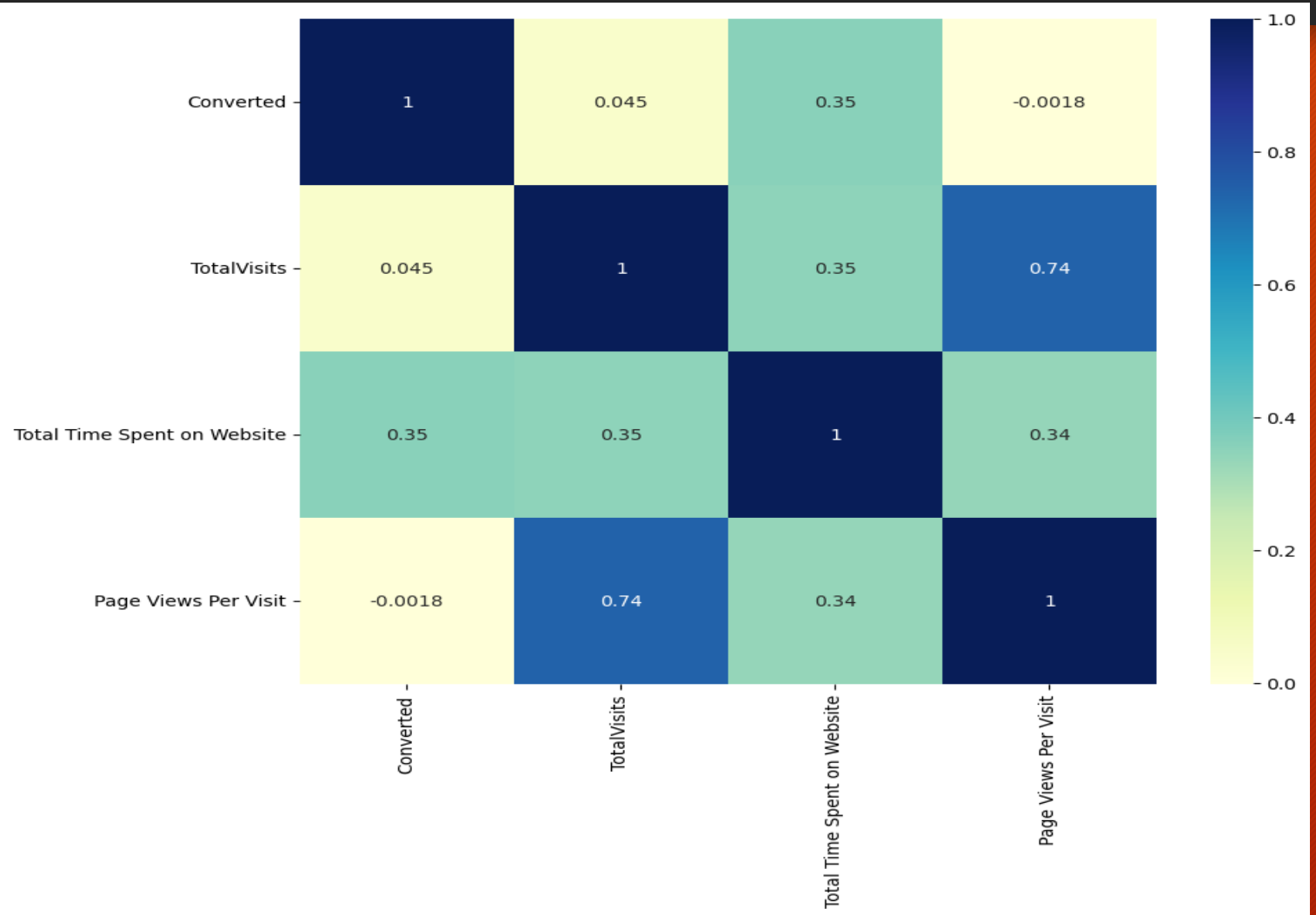

- **Outlier Treatment:**

Remove top & bottom 1% of the Column Outlier values

-

# Numerical Variable Analysis






finding the correlation among the numeric variables

# Dummy Categorical Variables



Creating dummy categorical variables to build the feature space

# Data Splitting



- Putting feature variable to X
- Putting target variable to y

- Splitting the data into train and test in the ration 70:30

# Model Building



- A model is built using stats model and RFE.
- Feature selection using RFE.
- For high VIF values of feature variables the column is dropped

**# confusion matrix**

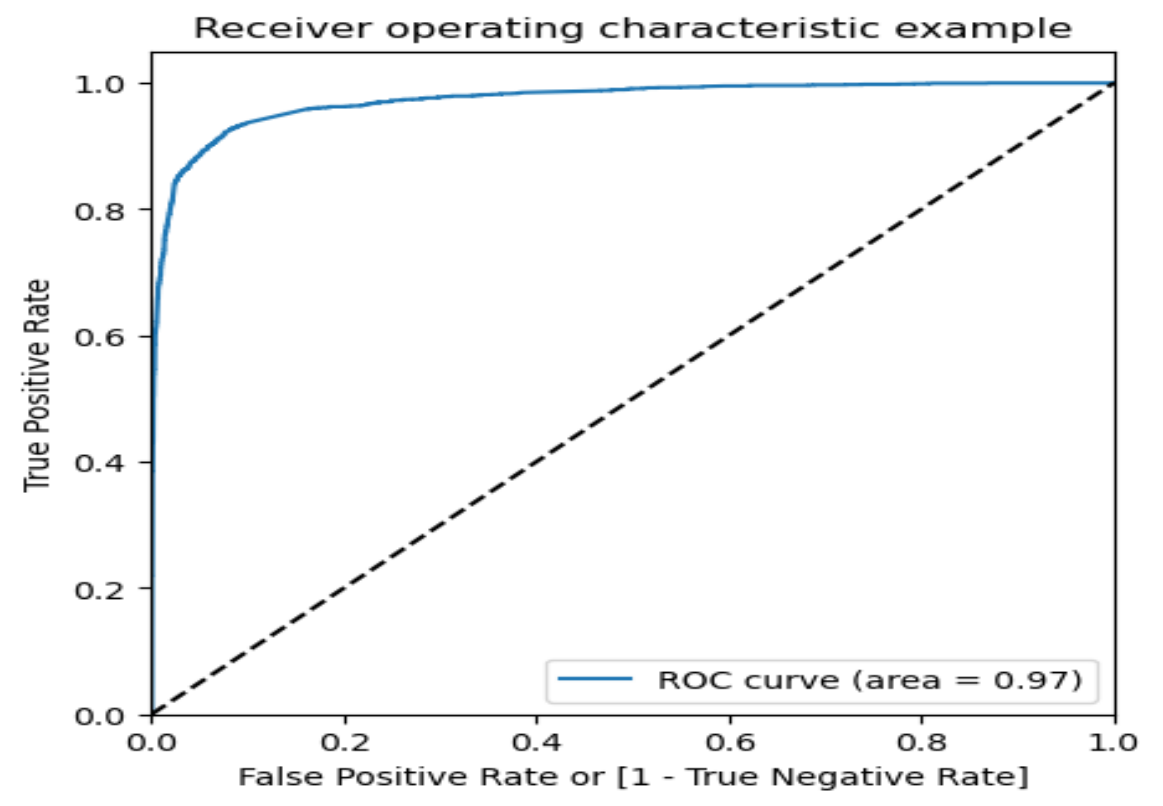
**[3774, 163],**

**[ 299, 2078]]**

# Model Building



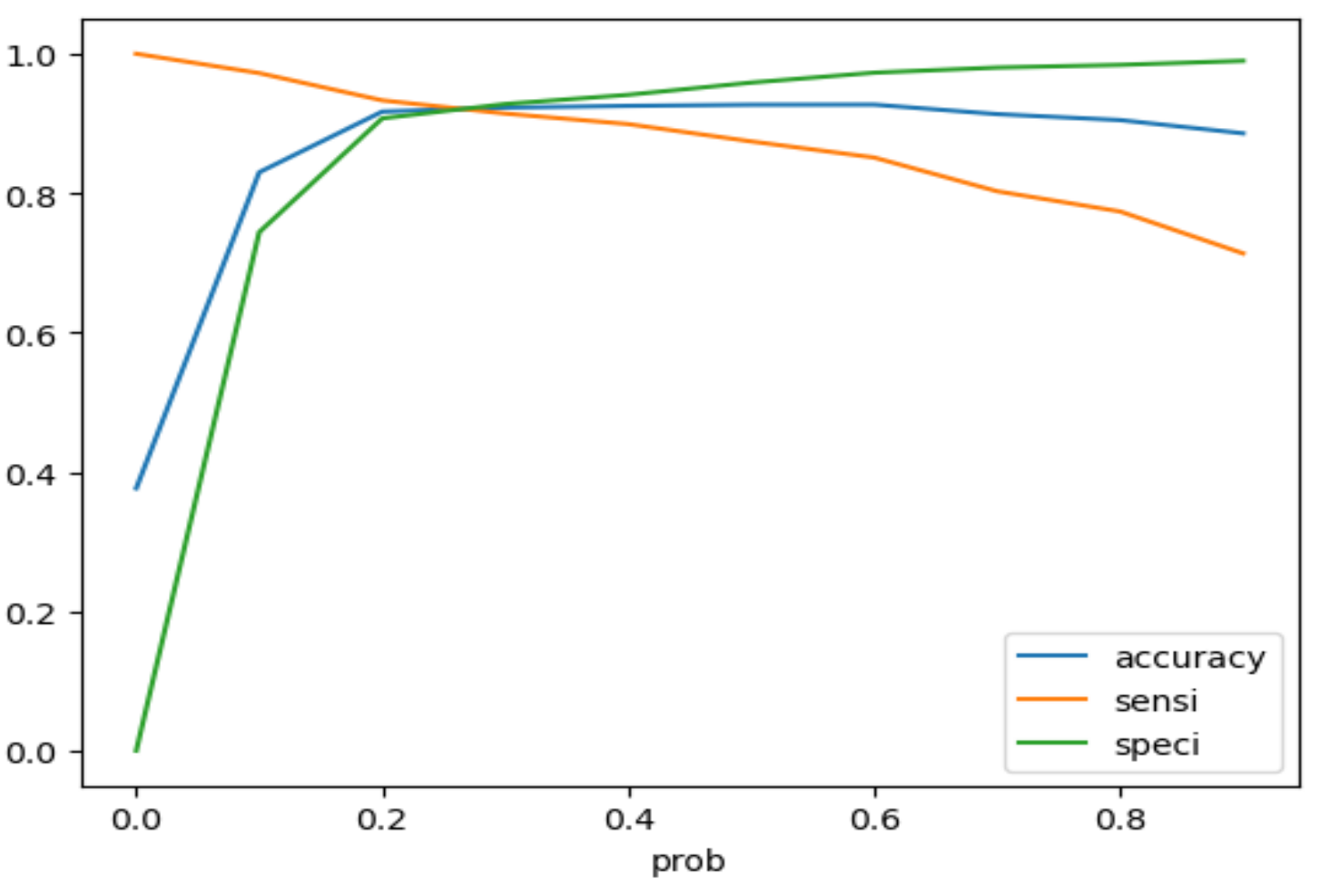
- Plot the ROC curve.




**The ROC Curve close to 1 is always bettr. We are getting a value of 0.97 indicates goot result**

# Model Building






**plot of accuracy, sensitivity and specificity for various probabilities shows the utt off of 0.3**

# Model Building



- The train data was found to have an
  - accuracy of 92.27%,
  - a sensitivity of 91.41%, and
  - a specificity of 88.44%.
- The test data had an
  - accuracy of 92.27%, a
  - sensitivity of 89.88%, and a
  - specificity of 90.14%.
- The model was found to predict the conversion rate well and the top features to consider were identified.

# Model Building: Testing samples analysis



confusion2 matrix

[3653, 284],

[ 204, 2173]],

It is observed that model is able predict true positive and true negative high percentage. And false negative is 4% which is quite low. Hence the model can be accepted

# Recommendations and conclusion



- **The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model**

**Top features to be considered which impact most in converting the lead to Enrolments**

- **What is your current occupation_Working Professional**
- **Total Time Spent on Website**
- **Lead Source_Olark Chat**
- **Lead Origin_Lead Add Form**
- **Last Activity_SMS Sent**
- **Lead Source_Welingak Website**
- **Tags_Will revert after reading the email**
- **Tags_Lost to EINS**
- **Tags_Closed by Horizzon**

# Recommendations and conclusion



- Since the accuracy of model designed is 92.27%, the findings are much more useful for considerations
  - The education institute has to focus more on working professionals
  - Has to create more engaging content on website
  - Need to send the SMS/message regularly
  - Get the feedback from email regularly

# THANK YOU