LEADS SCORING CASE STUDY

Group Members

1. Shivansh Bhardwaj
2. Shiva Ramkrishna
3. Shivkant Birthriya

Problem Statement

- X Education, an online course provider, generates leads through various channels such as their website, referrals, and platforms like Google. However, only about 30% of these leads currently convert into paying customers, which is a relatively low success rate.
- To enhance efficiency and maximize their efforts, the company aims to identify "Hot Leads"—high-potential prospects who are more likely to convert. This will allow the sales team to focus on engaging and nurturing these leads instead of contacting all prospects indiscriminately.
- Your role is to develop a model that assigns a lead score to each prospect, helping prioritize leads based on their likelihood to convert. The ultimate goal is to increase the lead conversion rate to approximately 80%.

Goals and Objectives

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution Methodology

1. Data Cleaning and Manipulation:

- Handle duplicate data and missing (NA) values.
- Drop columns with excessive missing values if irrelevant for analysis.
- Impute missing values when necessary and handle outliers.

2. Exploratory Data Analysis (EDA):

- Perform univariate analysis (value counts, variable distribution).
- Conduct bivariate analysis (correlation, variable patterns).

3. Feature Engineering:

- Apply feature scaling and create dummy variables for encoding.

4. Modeling and Validation:

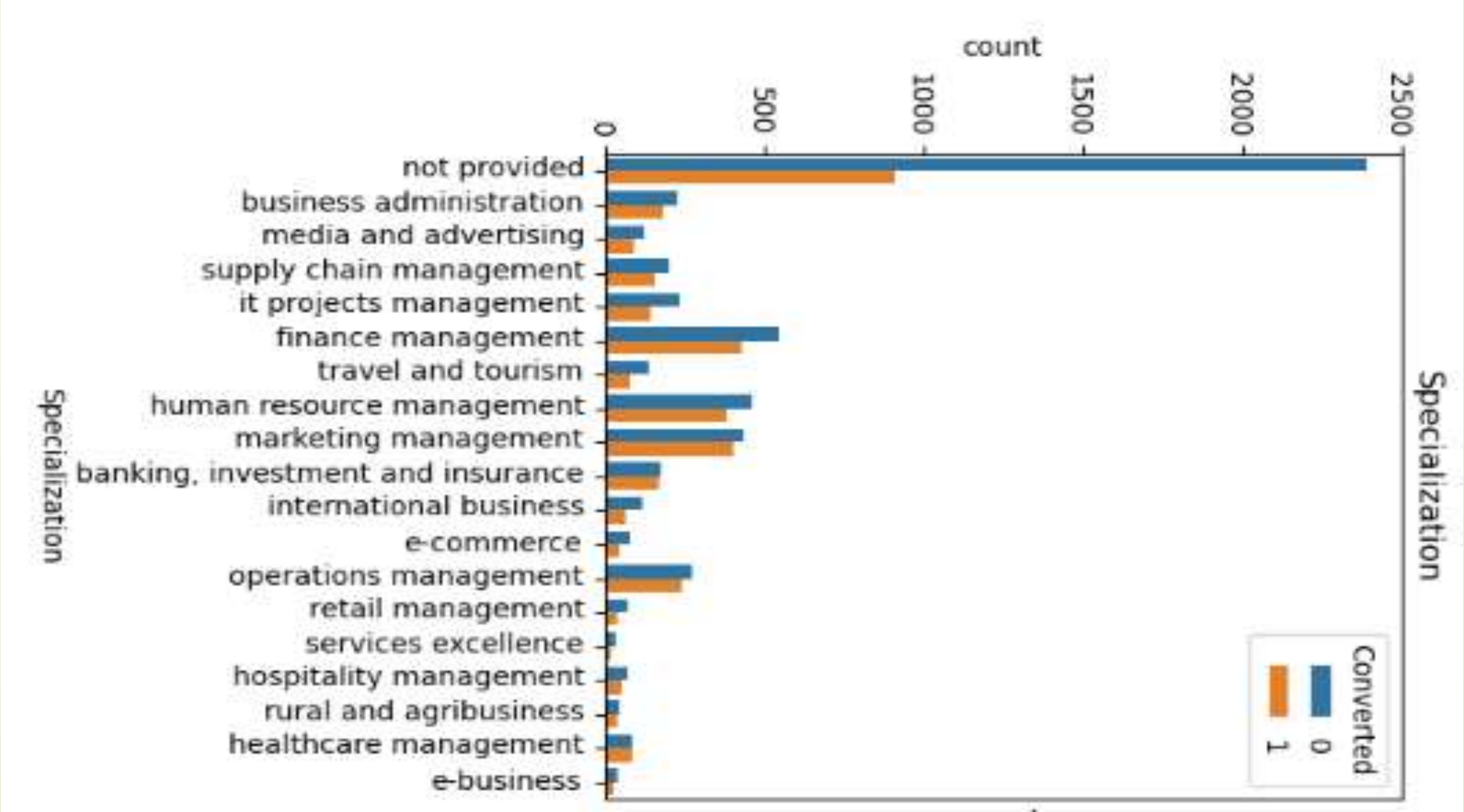
- Build and predict using logistic regression for classification.
- Validate model performance and present results.
- Draw conclusions and provide recommendations.

Data Manipulation

- The dataset has 37 rows and 9240 columns, with many irrelevant features.
- Features with single values, such as “Magazine” and “I agree to pay the amount through cheque,” were dropped as they don’t add value to the analysis.
- Columns like "Prospect ID" and "Lead Number," which don’t provide insights, were removed.
- Some object-type features, including “Do Not Call” and “What matters most to you in choosing course,” were dropped due to low variance and limited information.
- Columns with more than 35% missing data, like ‘How did you hear about X Education’ and ‘Lead Profile,’ were also removed to improve dataset quality.

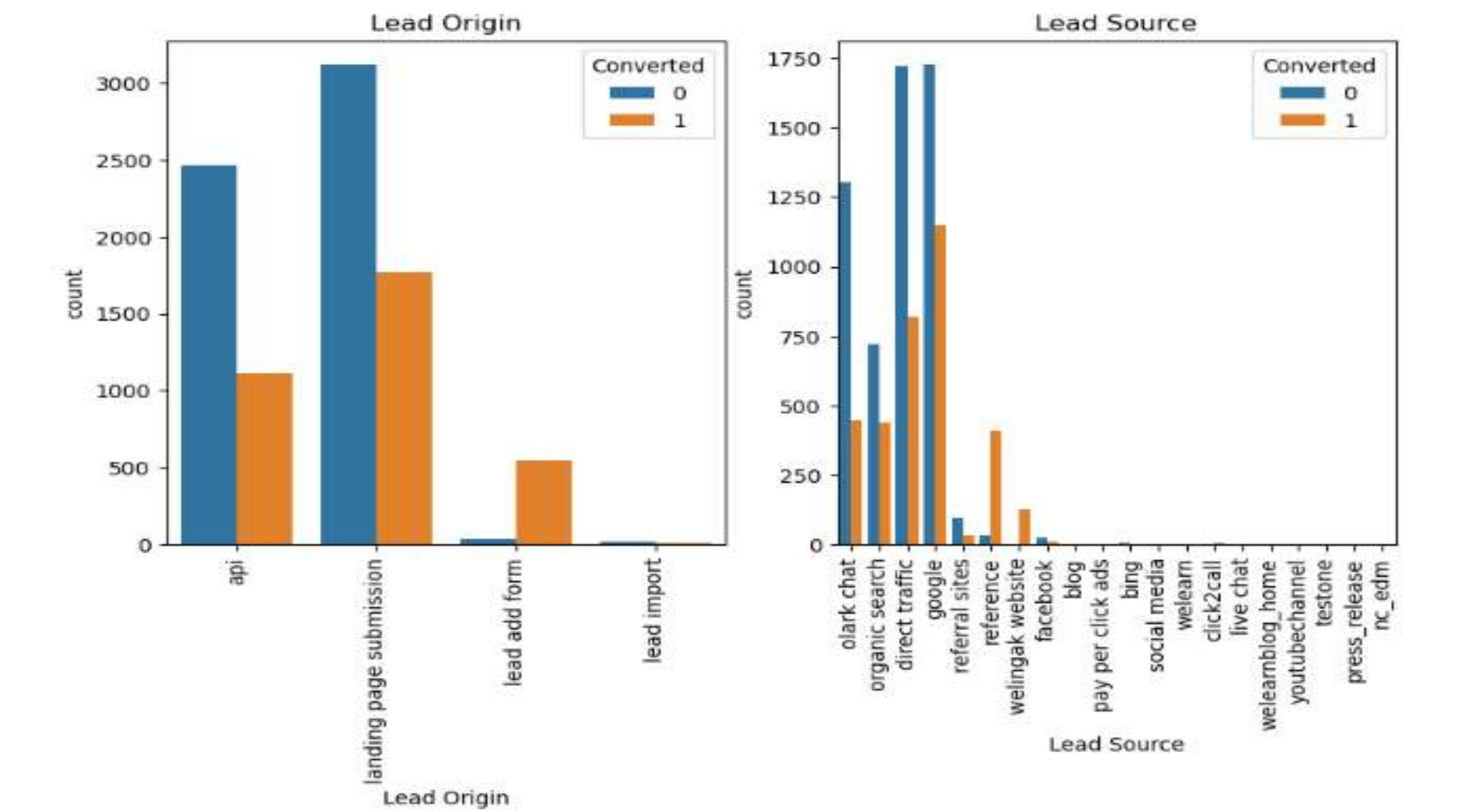
Specialization

Leads from HR, Finance & Marketing management specializations are high probability to convert.



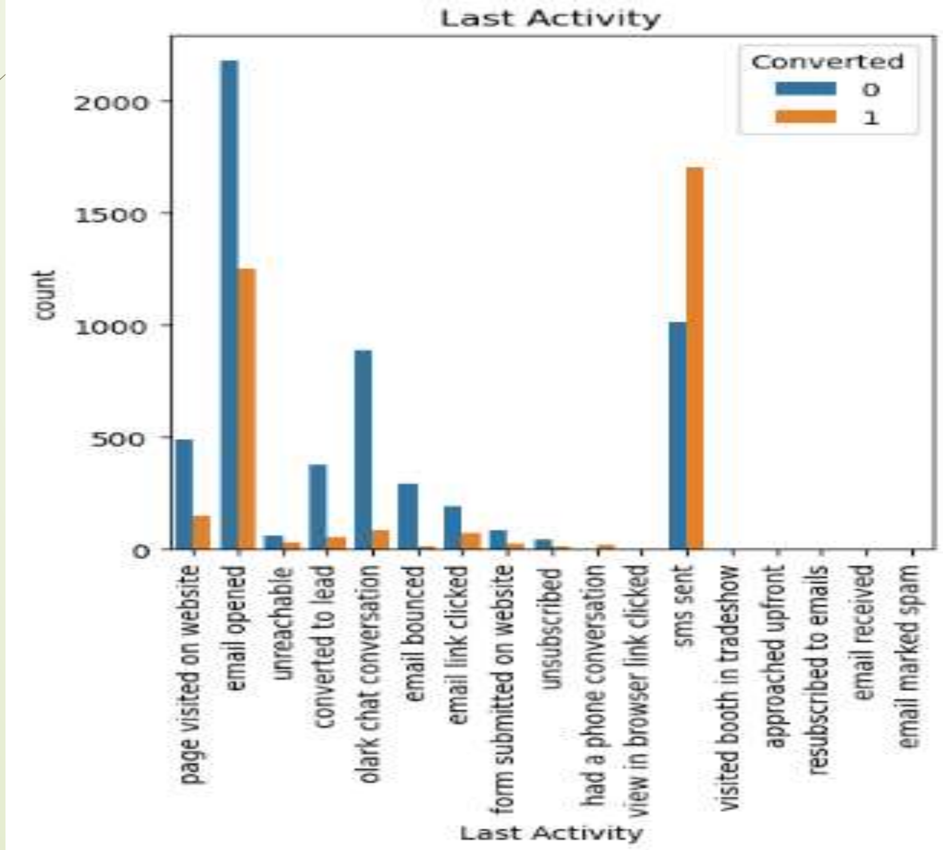
Lead Source & Lead Origin

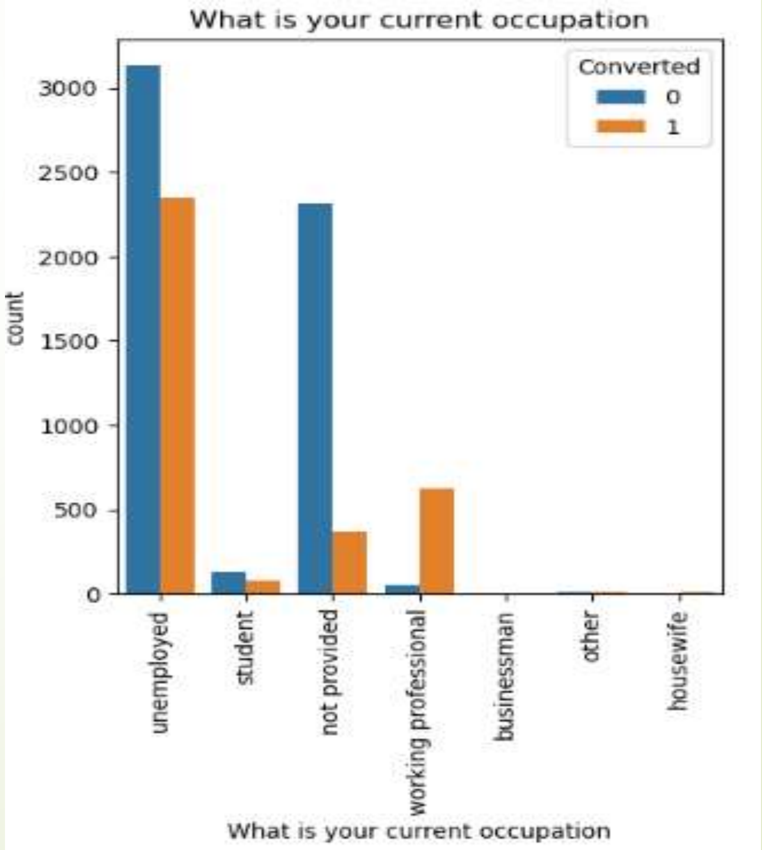
In lead source the leads through google & direct traffic high probability to convert.Whereas in Lead origin most number of leads are landing on submission.



Last lead Activity & What is Your Current Occupation

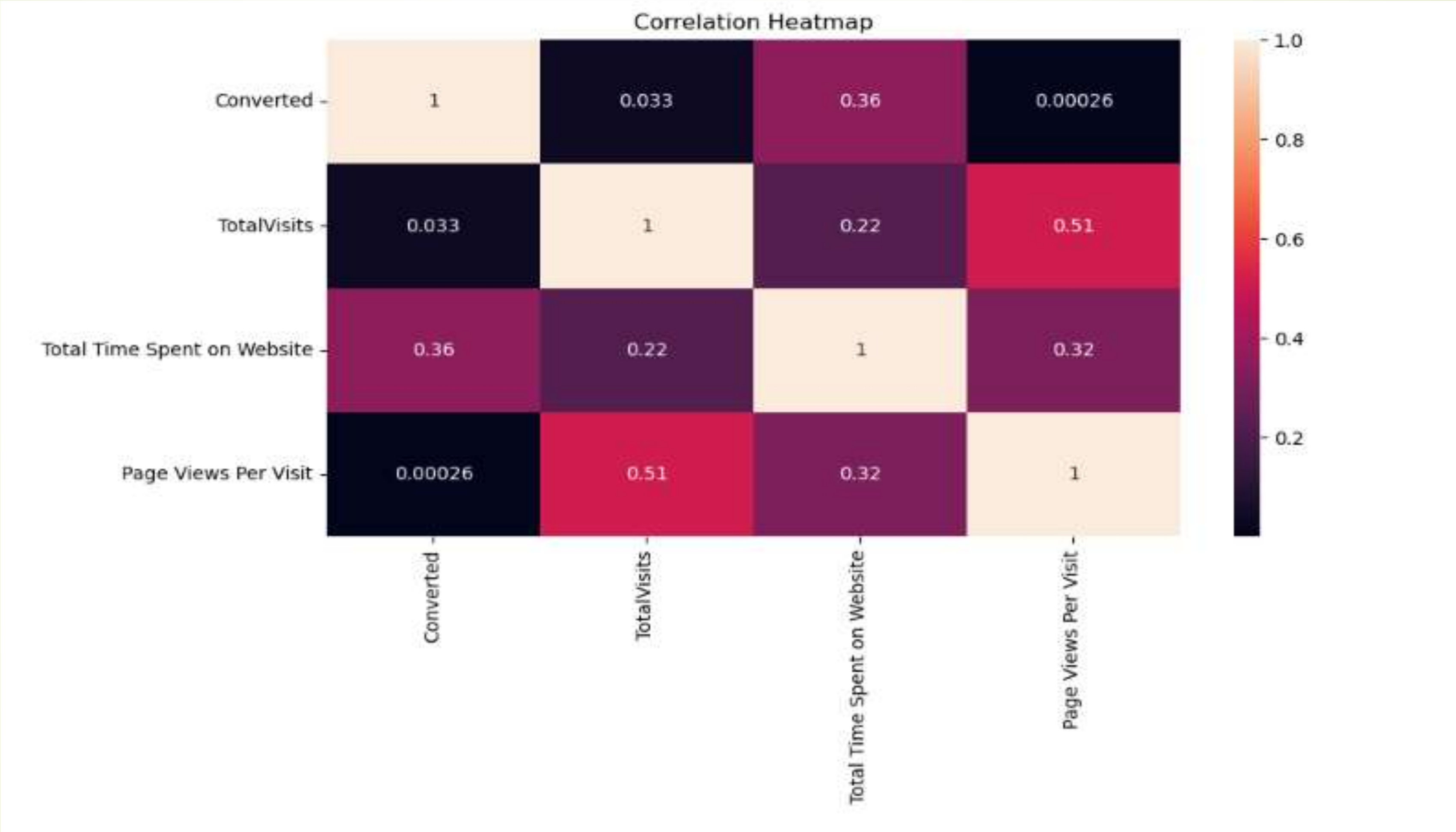
Leads opening emails or receiving SMS have a higher conversion probability, and unemployed leads are more interested in joining the course than others.





Correlation

There is no correlation between the variables.



Data Conversion

- Numerical variables have been normalised.
- Dummy variables have been created for categorical variables .
- Total number of rows for analysis: 8792
- Total number of columns for analysis: 4

Data Building

- Splitting the Data into Training and Testing Sets:

  The initial step for regression is splitting the data, and we've chosen a 70:30 ratio for training and testing.

- Using RFE for Feature Selection:

  RFE is applied with 15 variables as the output for feature selection.

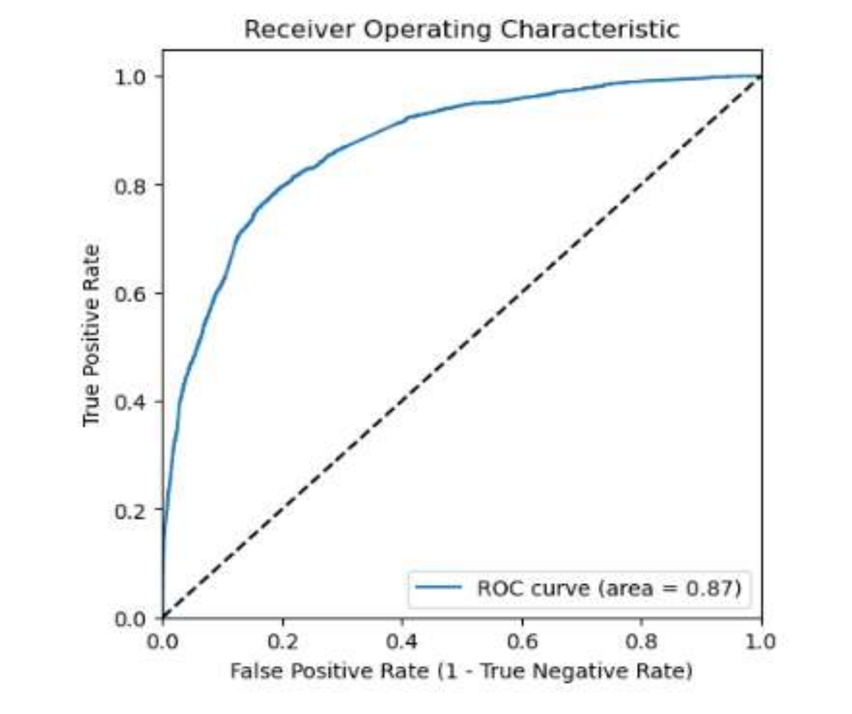
- Building the Model:

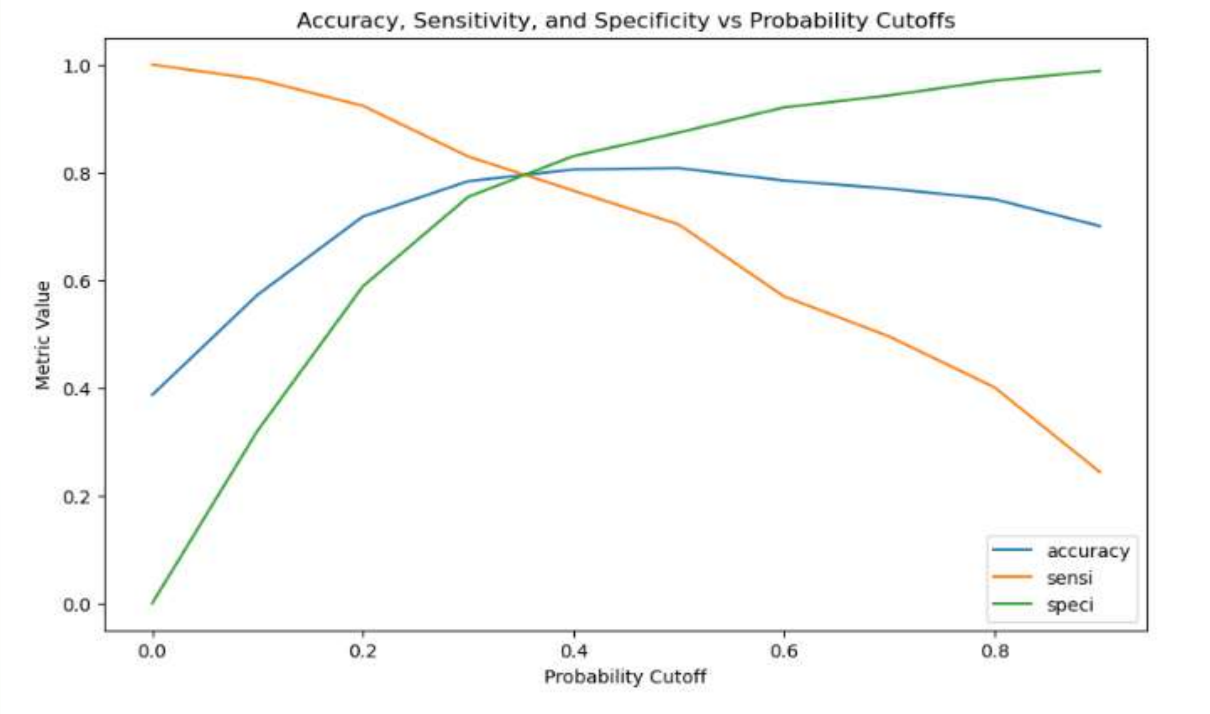
  Variables with a p-value greater than 0.05 and VIF value over 5 are removed to improve the model's performance.

- Predictions on Test Data:

  The model is evaluated on the test dataset, with an overall accuracy of 81%.

ROC Curve





Finding Optimal Cut off Point :

Optimal cut off probability is that probability where we get balanced sensitivity and specificity.

From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

The most important factors influencing potential buyers (in order of importance) are:

1. Total time spent on the website.
2. Total number of visits.
3. Lead source, specifically:
   - Google
   - Direct traffic
   - Organic search
   - Welingak website

Conclusion

4. Last activity, specifically:

    - SMS
    - Olark chat conversations

5. Lead origin, particularly when it is "Lead Add Format" .

6. Current occupation, specifically if the lead is a working professional.

By focusing on these key factors, X Education has a high potential to convert nearly all potential buyers into paying customers, significantly boosting enrollment in their courses.