Lead Score Case Study

Group Members
1. Pradeep Shivaji Dekhane
2. Pooja Mendigeri
3. Pooja BV

Problem Statement

- X Education, which offers online courses for working professionals, acquires a large volume of leads daily but converts only 30% on average.
- To streamline their sales process and improve efficiency, the company seeks to identify "Hot Leads" – leads with a high probability of conversion.
- By focusing sales efforts on these potential customers, X Education expects to significantly increase their overall lead conversion rate.

Business Objective:

- X education's objective is to maximize sales conversions.
- For that they want to build a Model by identifying and targeting "hot leads“ using a predictive model deployed for ongoing use.

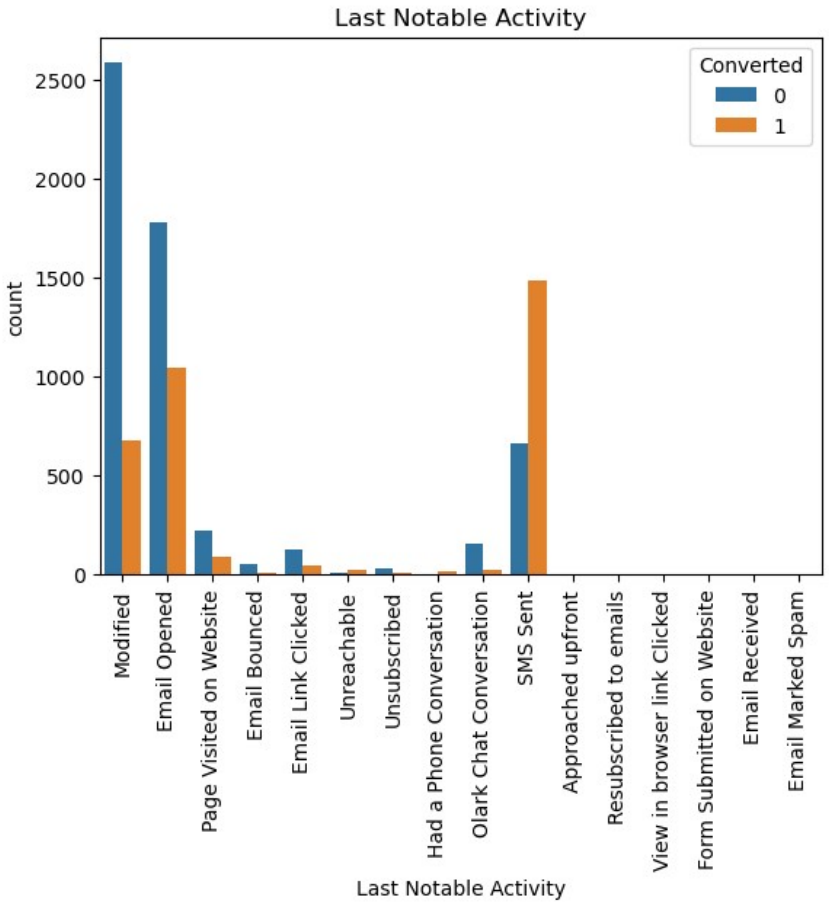
Solution Methodology

- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- Exploratory Data Analysis
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

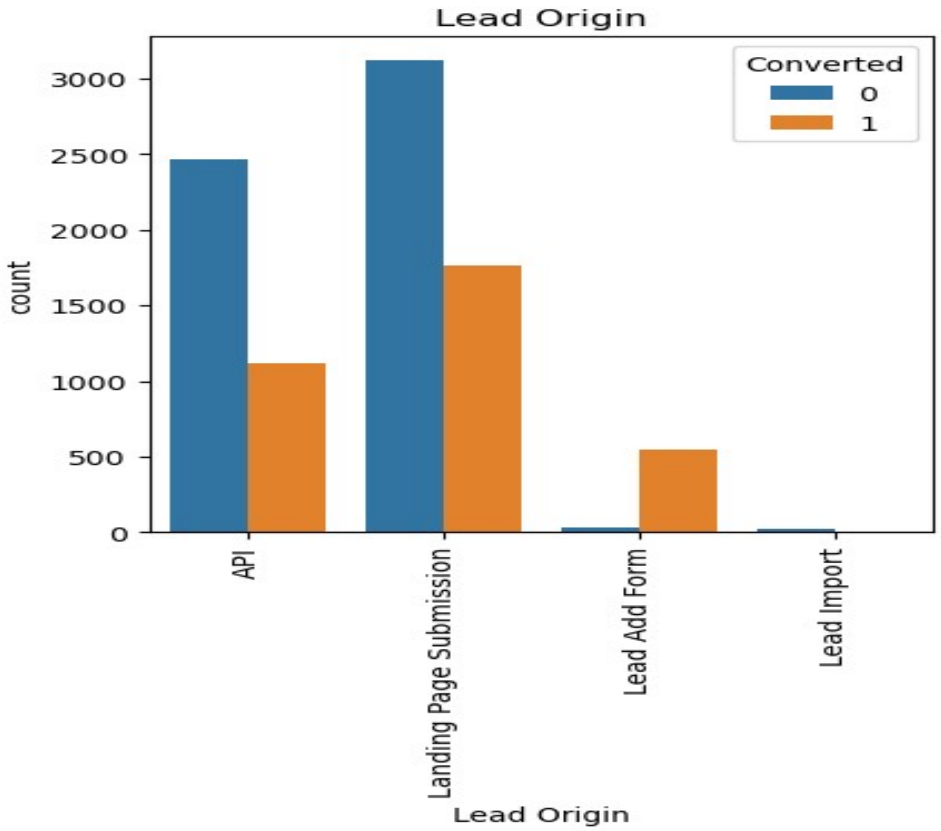
Data Manipulation

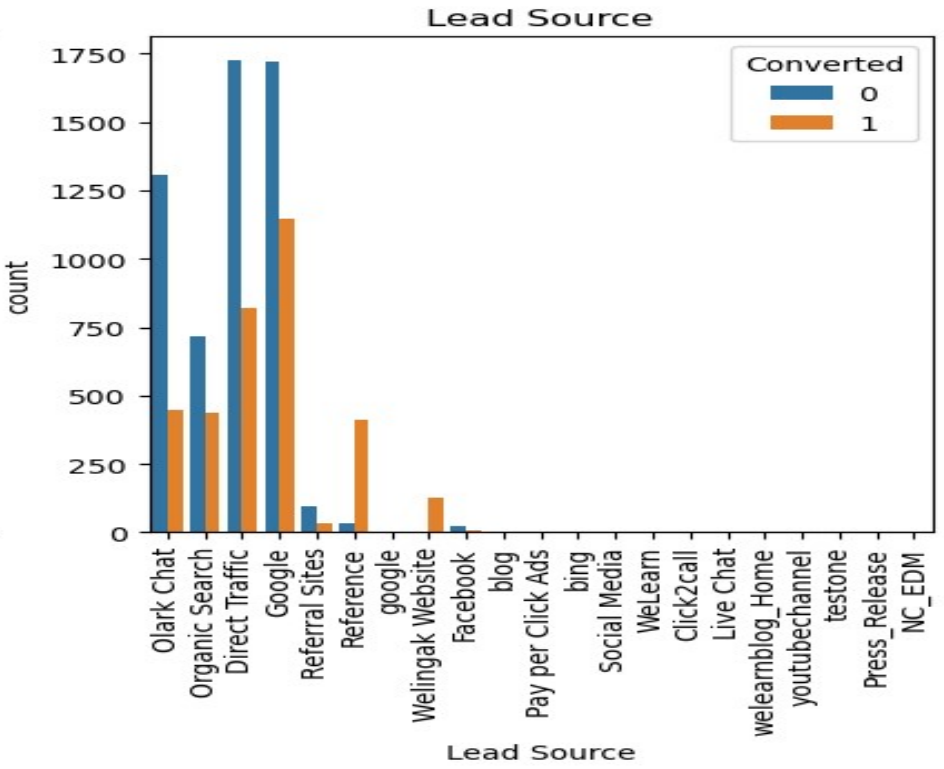
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Exploratory Data Analysis



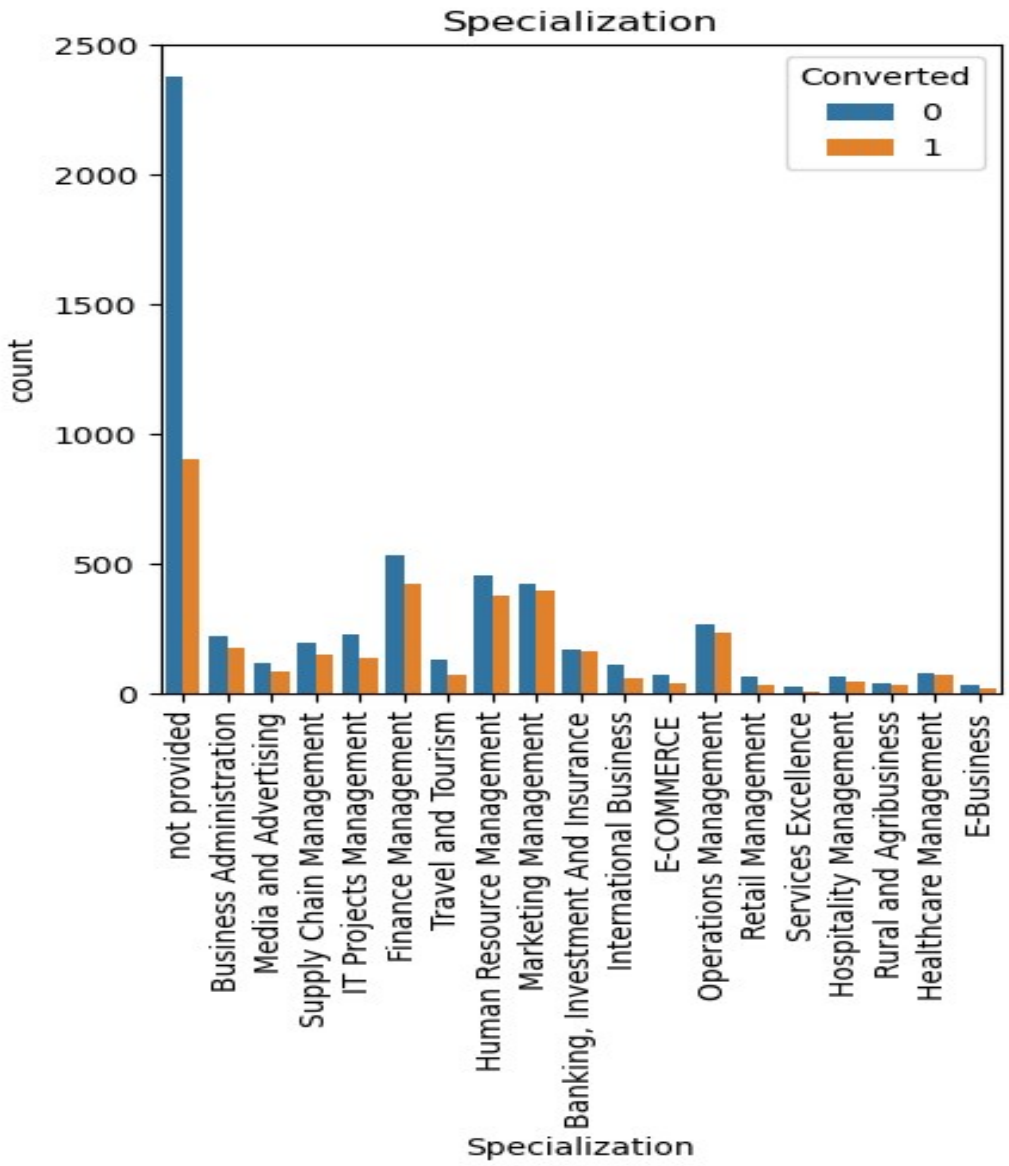
- The most frequent "Last Notable Activity" is "Modified," with a significantly higher count than any other activity.
- "Email Opened" has a noticeable difference, with more opens for converted users. "Email Bounced" and "Email Marked Spam" have very low counts overall.
- The Converted users is slightly taller than the non-converted for "SMS Sent." This *suggests* that users whose last notable activity was an SMS might have a slightly higher chance of converting.

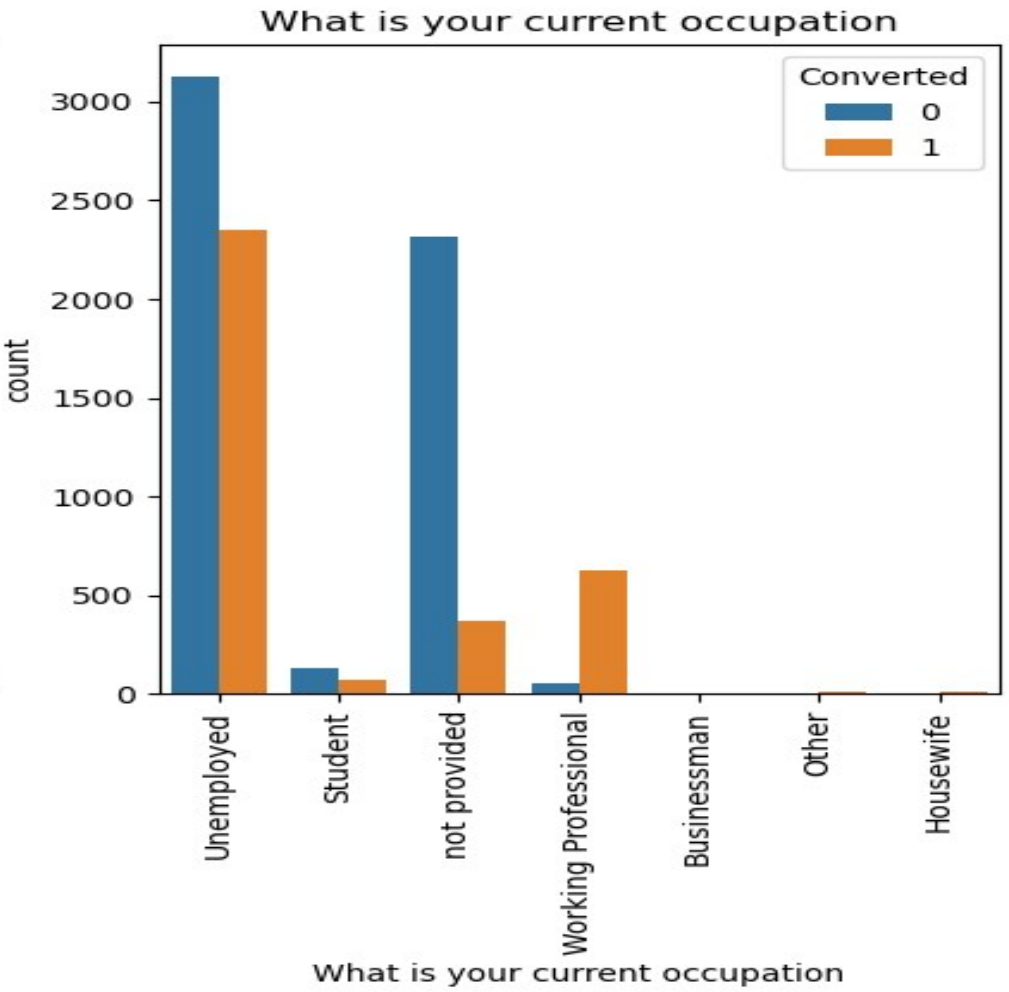


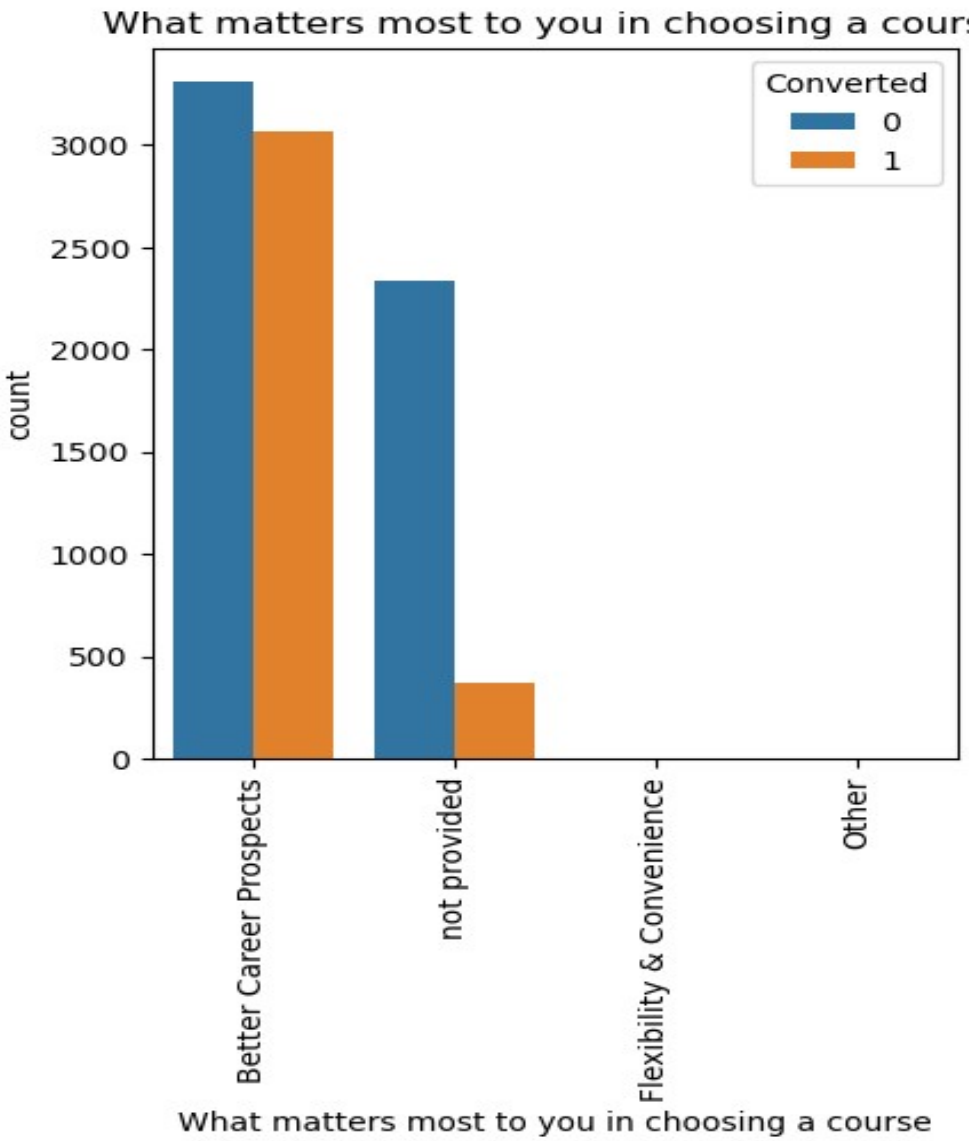


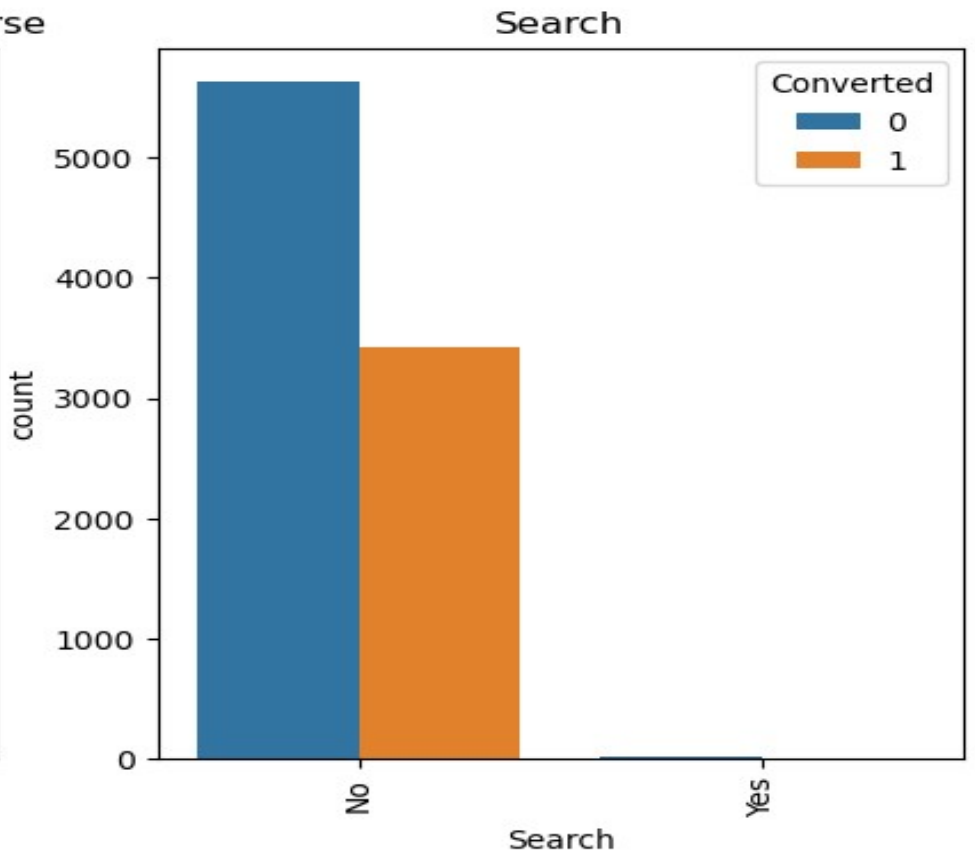
Landing page submission in “Lead Origin” has high lead conversions

Google searches in “Lead Source” has high conversions compared to other modes, while references has high conversion rate.


Specialization
Converted
0
1
count
2500
2000
1500
1000
500
0
not provided
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
not provided
Working Professional
Businessman
Other
Housewife
What is your current occupation



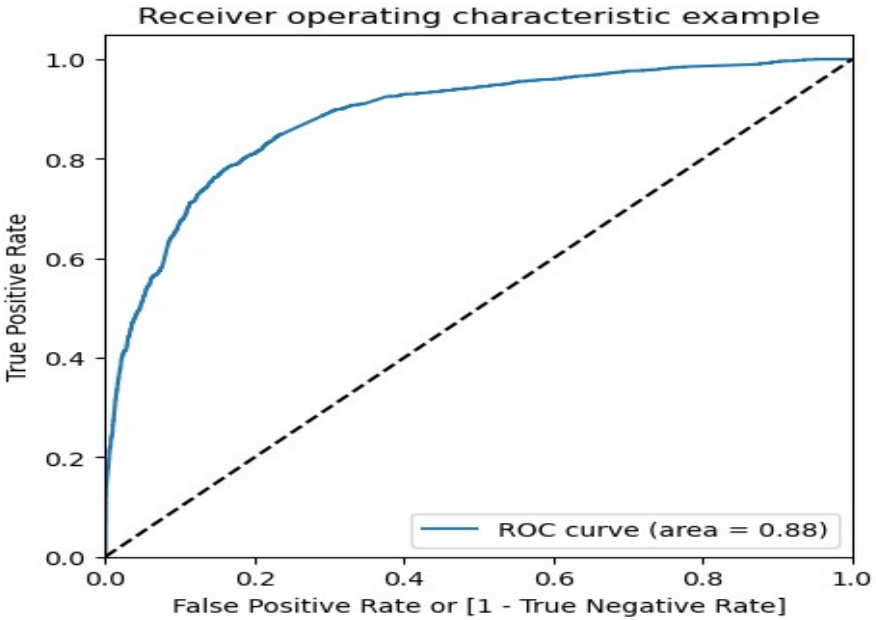


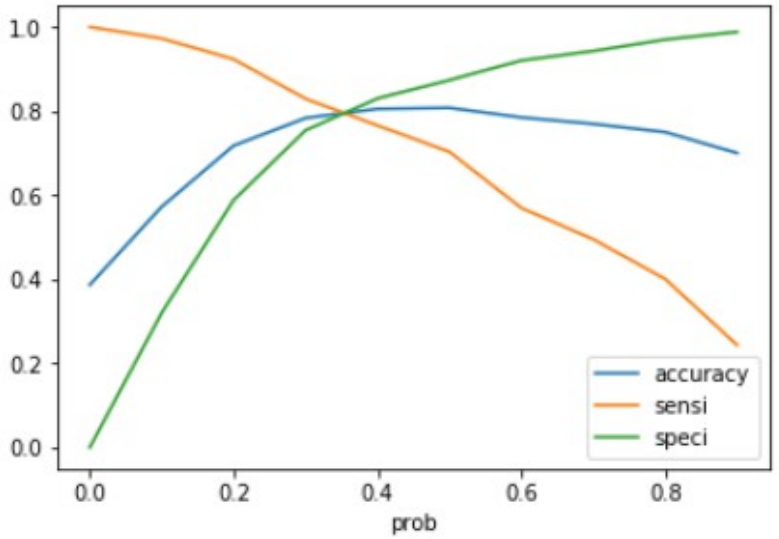
- The above graph shows searches are not good leads.
- Better career prospects, the conversion rate is slightly *lower* proportionally compared to non-converted individuals. This suggests that while important, it's not the *sole* driver of conversion.

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

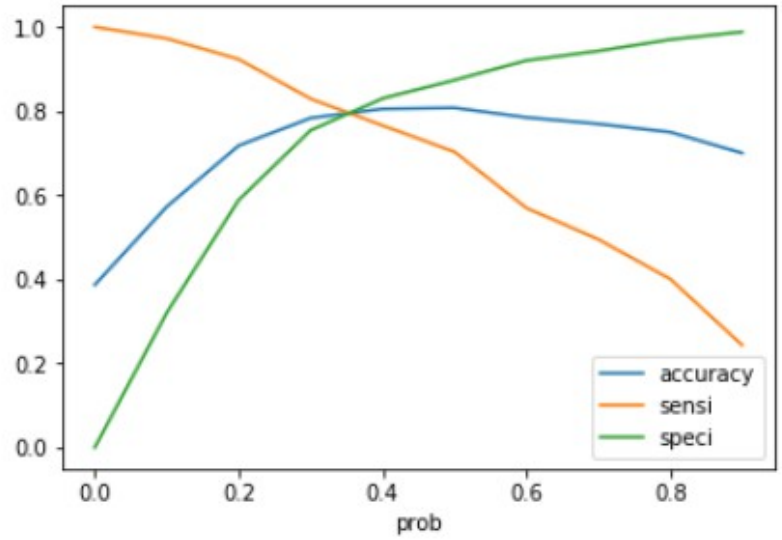
ROC Curve



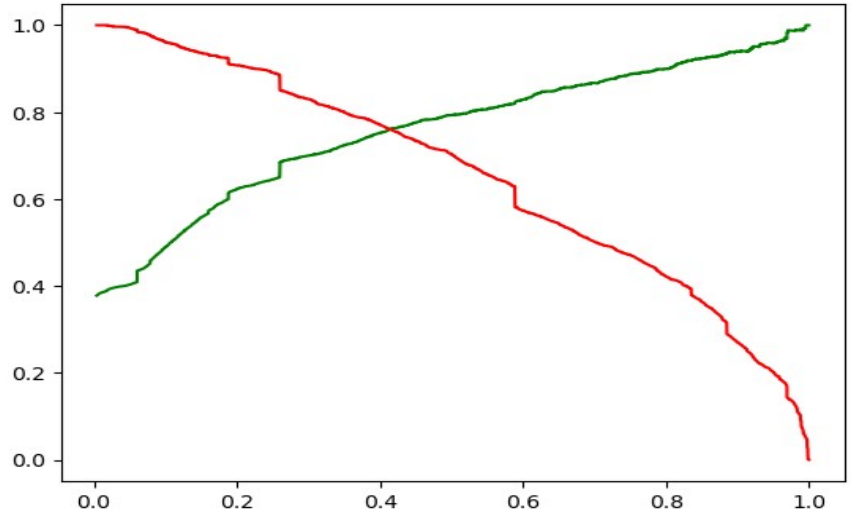


- **Finding Optimal Cut off Point**
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

Model Evaluation Train and Test



- Accuracy Sensitivity and Specificity is 81%, 80% and 82% respectively.
- Precision and Recall is 79% and 70% respectively.
- From the above graph it is visible that the optimal cut off is at 0.35.



- Accuracy Sensitivity and Specificity is 80%, 79% and 81% respectively.
- Precision and Recall is 75% and 76% respectively.
- From the above graph it is visible that the optimal cut off is at 0.41.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.
- Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.