Lead scoring case study

By Amarjot Singh, Debasish Rath and Sanjan Shetty

Problem statement

- X Education is an organization which provides online courses for industry professionals
- The company generates a lot of leads, but only a few of them are converted into paying customer. Also the company wants a higher lead conversion. It can be done through Google searches, online advertising, paid promotions, various online platforms etc
- The conversion rate is approx. 30% out of 100% and is very poor in implementing the process of generating the customer to enroll into the online studies.
- They need to find a way to make the conversion rate >80% to become most promising leads. For that they need to build a model through which the issue can be identified and can later on resolved for the better future

Strategy

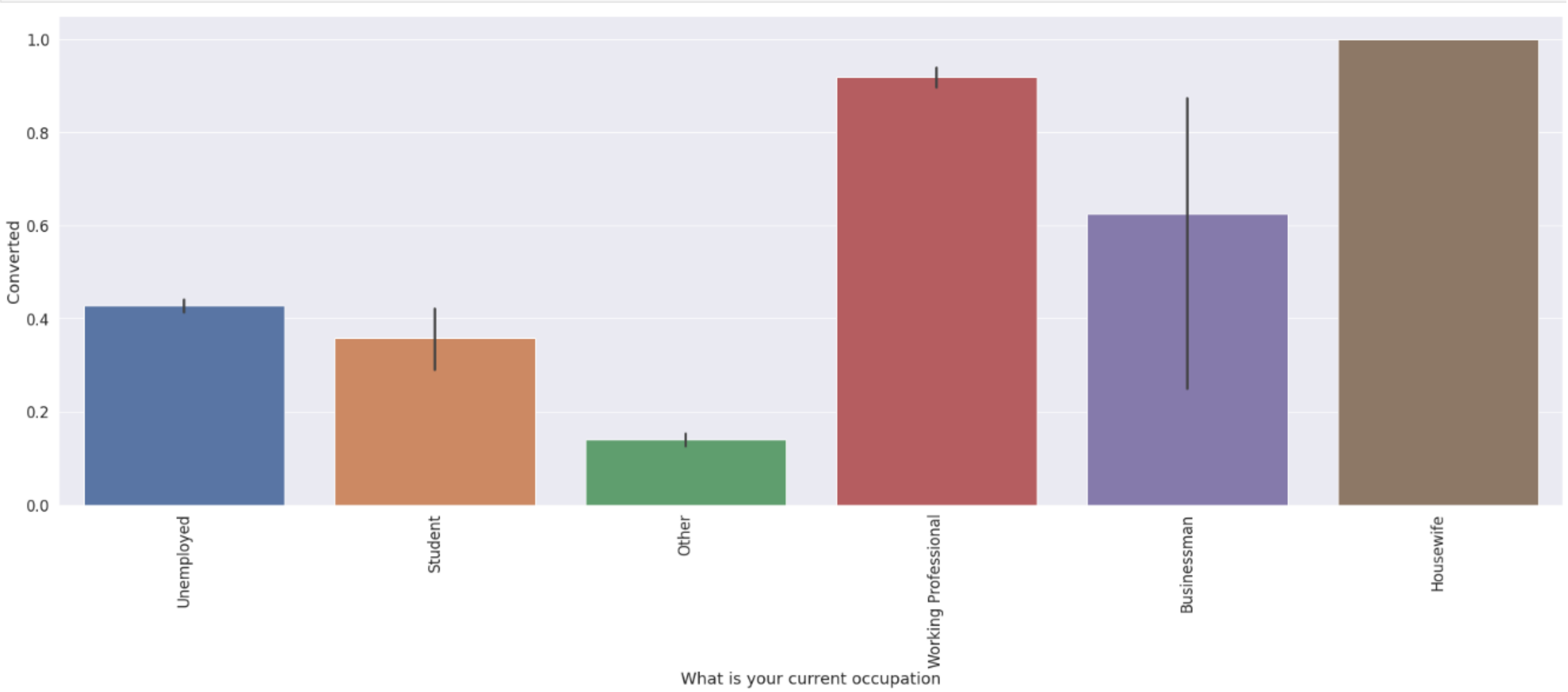
- Reading the dataset
- Inspecting
- Cleaning the null values:
  - Handling the duplicate values
  - Dropping of the columns which are unnecessary or having large amount of missing values
  - Imputation of data, handling outliers in data
- EDA:
  - Approach to analyze the data using visual techniques.
  - Univariate and Bivariate data analysis
- Transformation (Get dummies/External columns)
- Train test split:
  - Min and Max scaler
- Machine learning model-Logistic regression:
  - Results
  - Variable
  - P value, VIF corr()

Inspecting and cleaning the null values

- In total there are 9,240 rows and 37 columns
- Chain Content, Get updates on DM Content, I agree to pay the amount through cheque etc. have been dropped
- Had to remove the Prospect ID and Lead Number which is not necessary for the analysis
- While scrutinizing the value counts for some of the object type variables, we found some of the features which has no enough variance, and decide to drop columns such as Do Not Call, What matters most to you in choosing course, Search, Newspaper Article, X Education Forums, etc.
- Some columns consist of single values features like Magazine, Receive More Updates About Our Courses etc
- Analyzed values by country wise, specialization etc

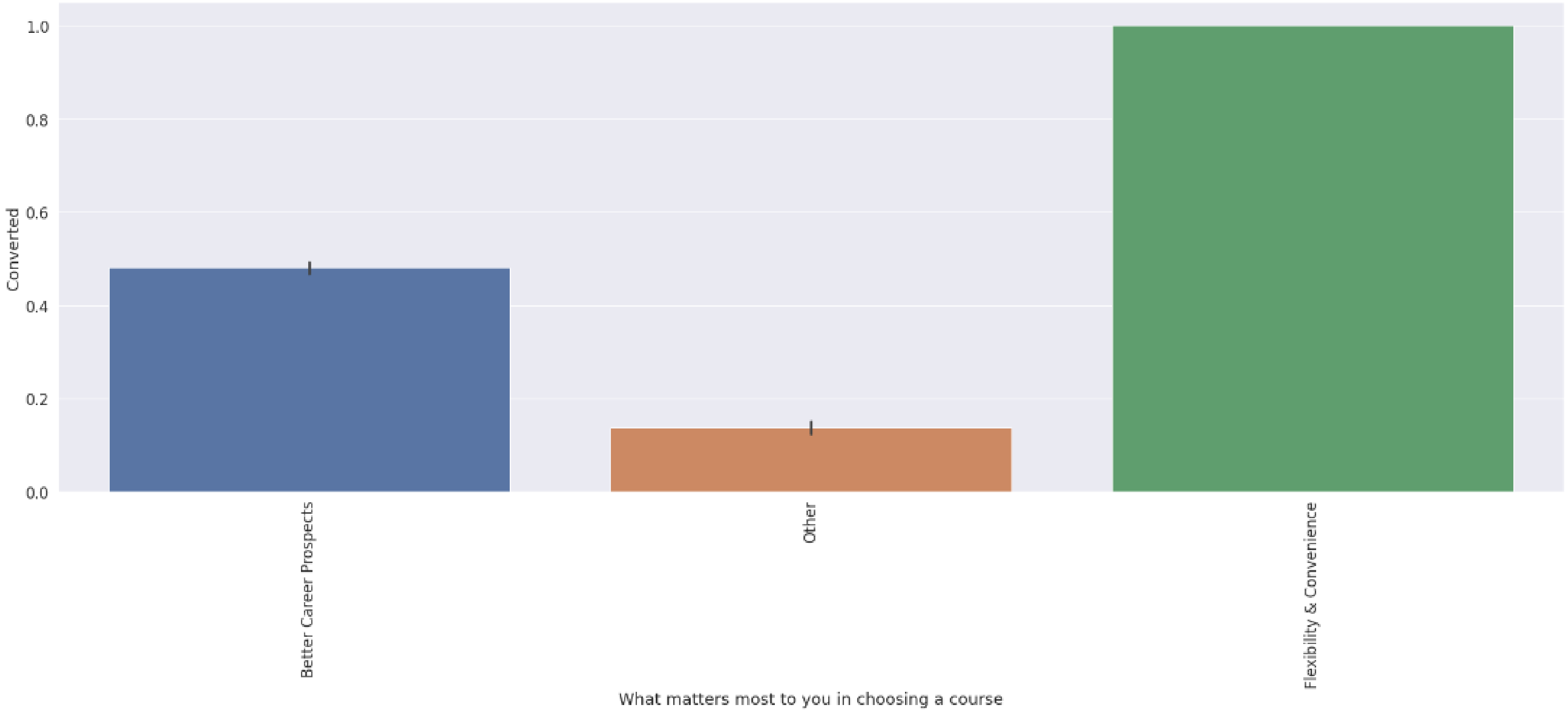
EDA

- With the help of the chart, we can identify how many people are Unemployed, students, working professional etc. who want to enroll into online courses

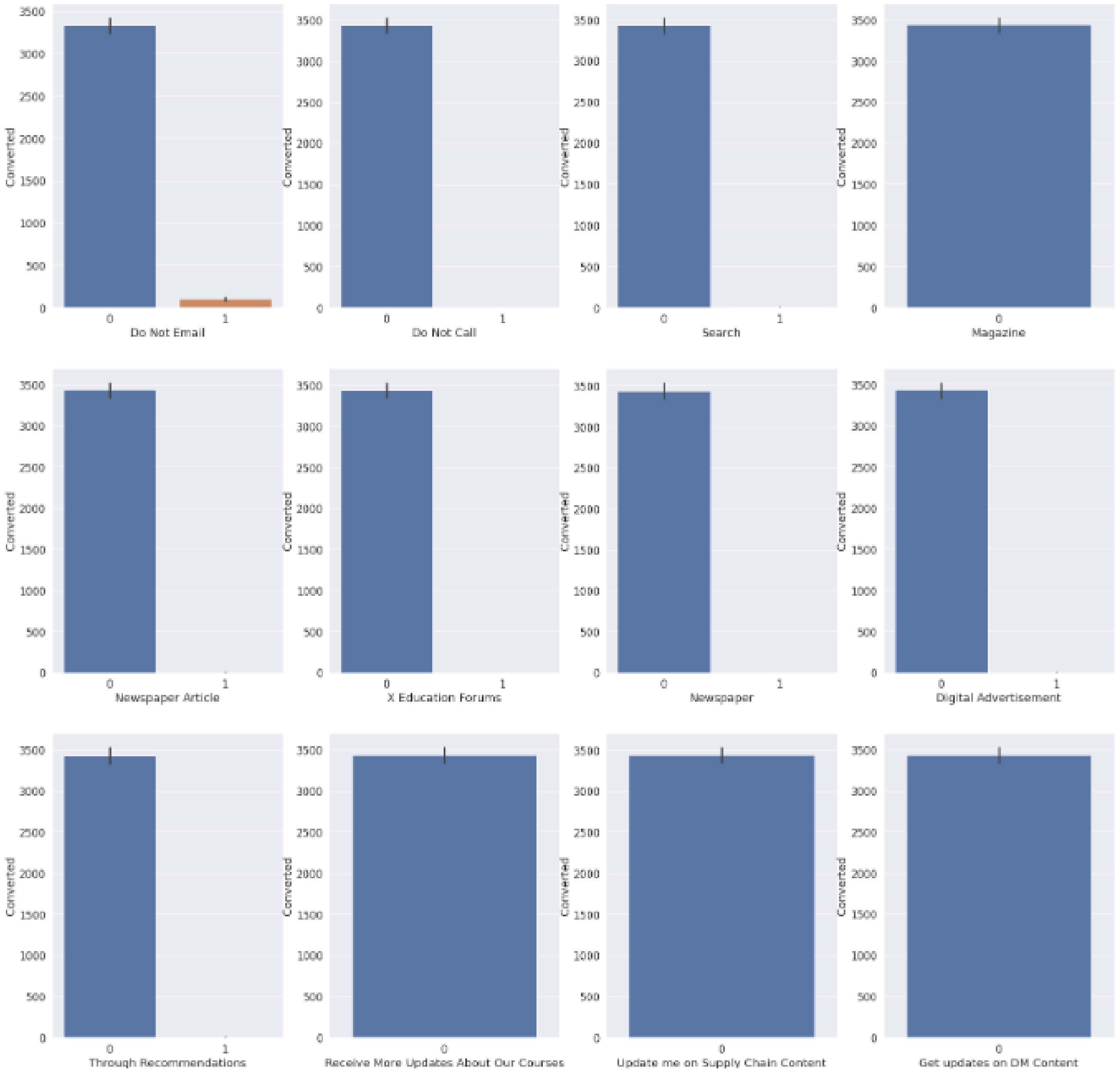


EDA (Contd.)

- Through this we can come to know what is the reasons people are preferring online courses

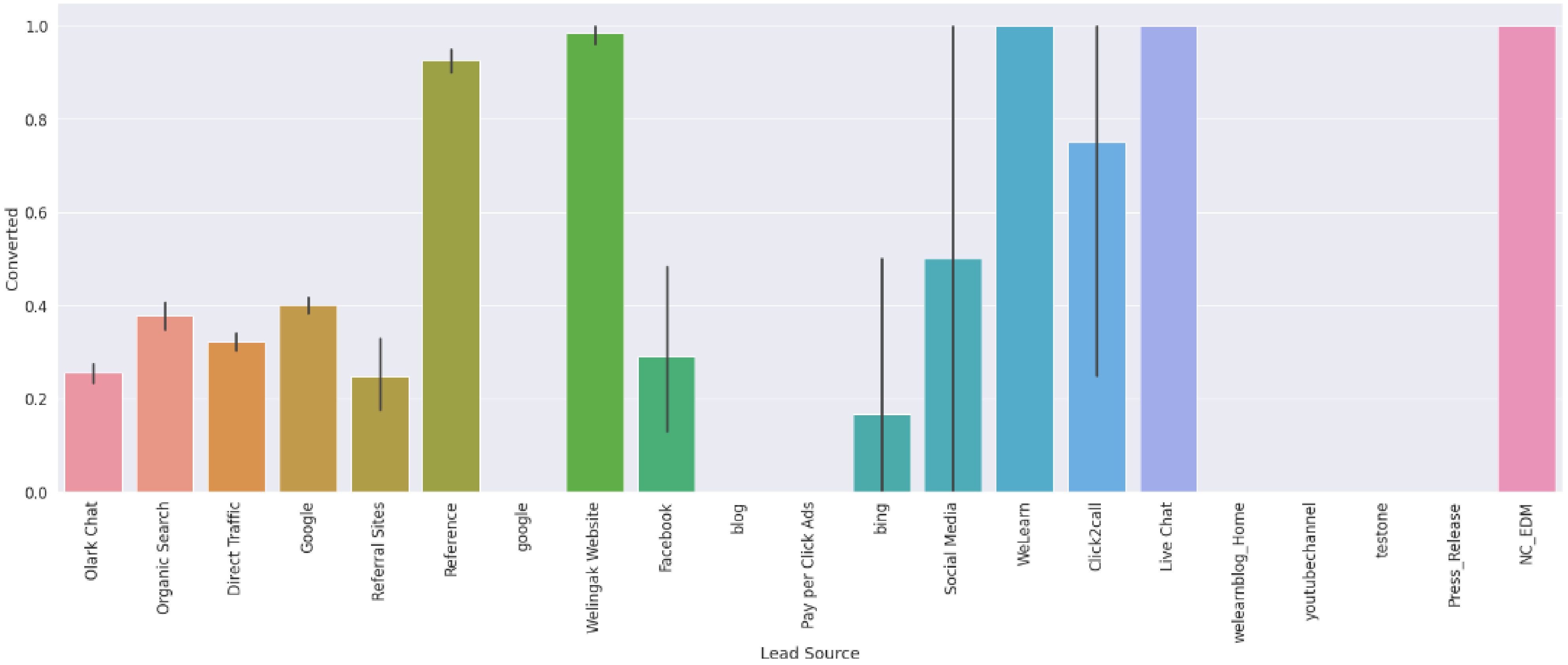


EDA (Contd.)


Converted
3500
3000
2500
2000
1500
1000
500
0
0
1
Do Not Email
Do Not Call
Search
Magazine
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendations
Receive More Updates About Our Courses
Update me on Supply Chain Content
Get updates on DM Content

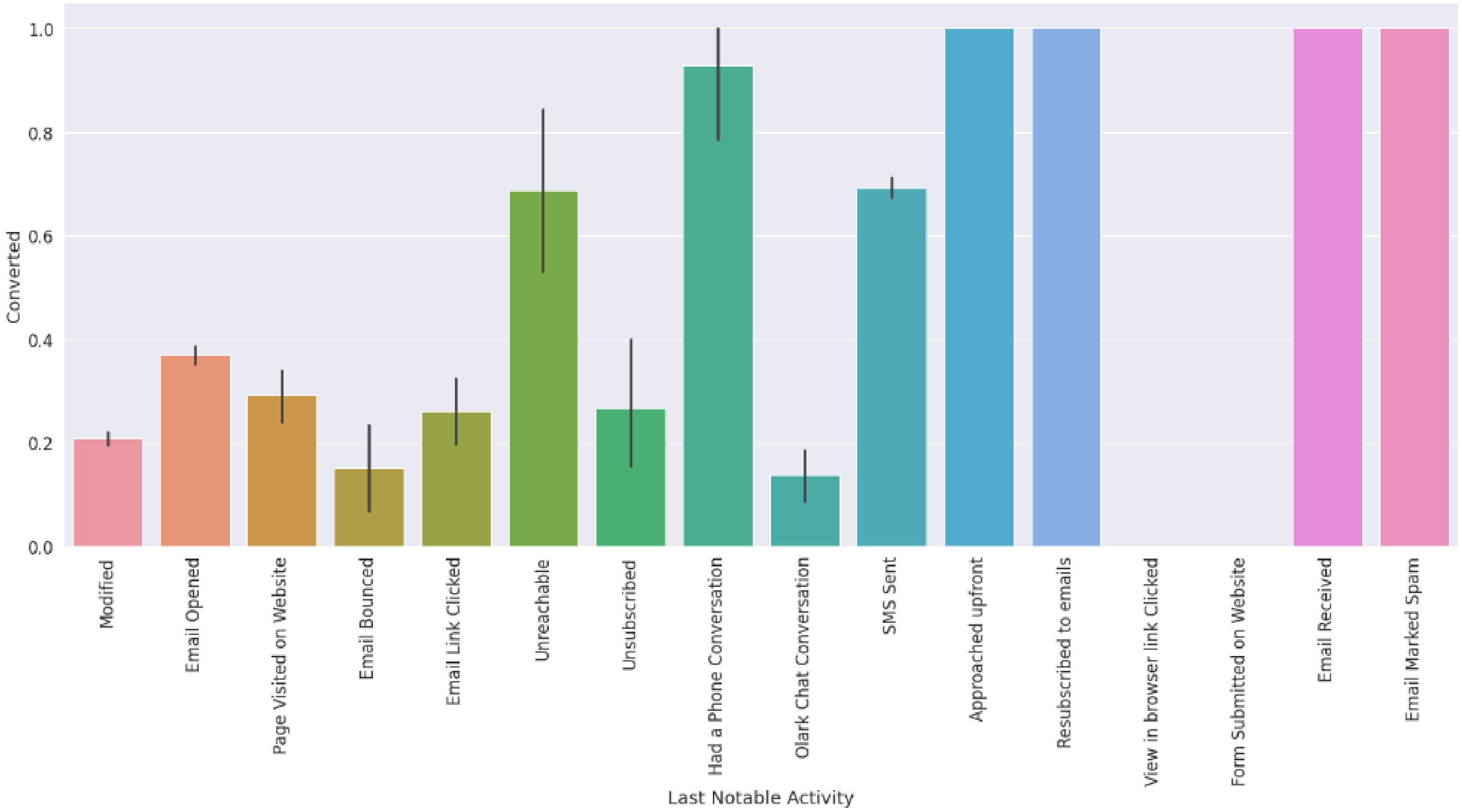
EDA (Contd.)

- With the help of the below chart, we can come through know how do they came to know about the online courses offered by the company



EDA (Contd.)

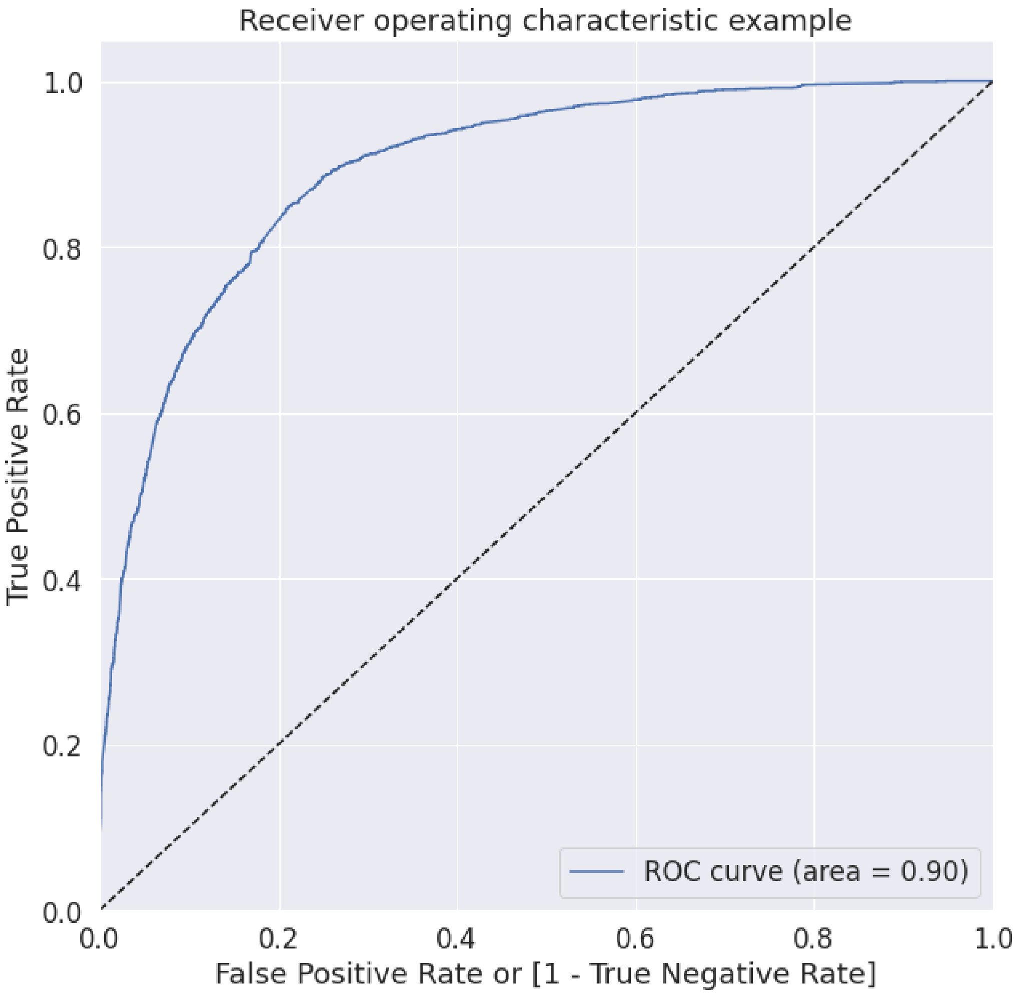
- The Clients Approach upfront, resubscribed to emails and received an email are most likely to convert



Model building

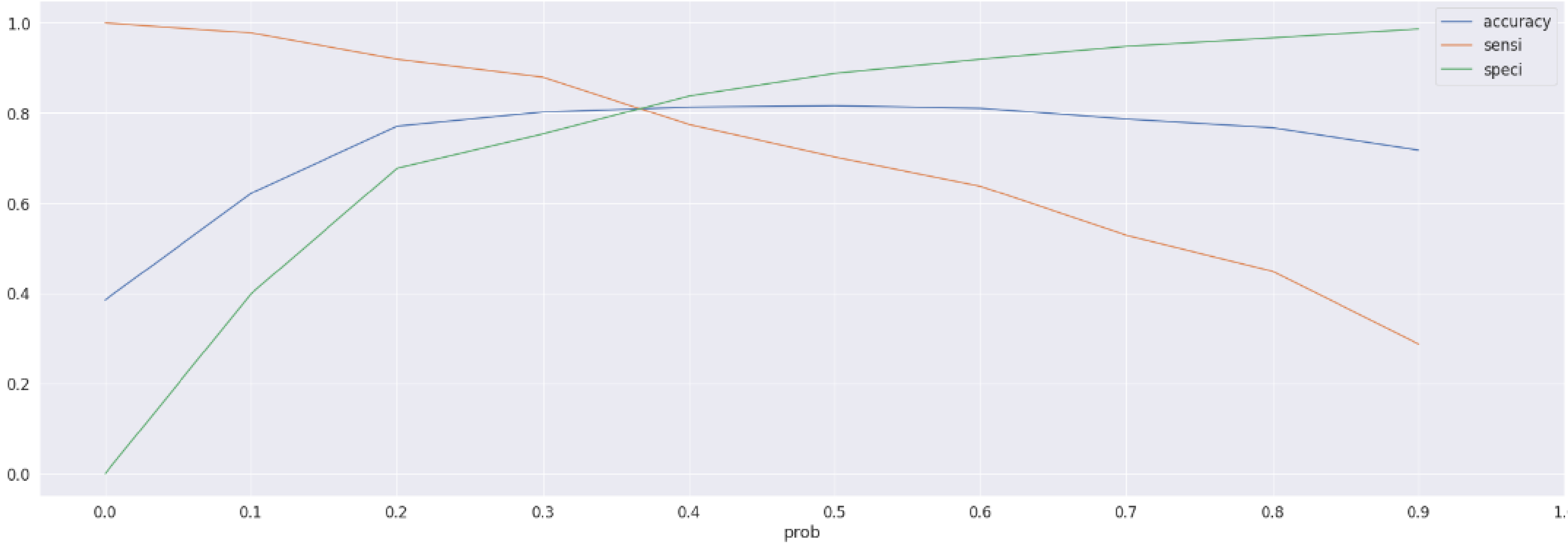
- Splitting the data into Train and Test sets
- For regression, need to perform train-test split. The split ratio should be 70-30
- Use RFE to eliminate the relevant variables
- Running RFE with 15 variables as output
- Build a Model by removing the variable, whose p- value is greater than 0.05 and VIF value is >5
- Predict using test set
- Precision and recall analysis on test predictions

ROC curve

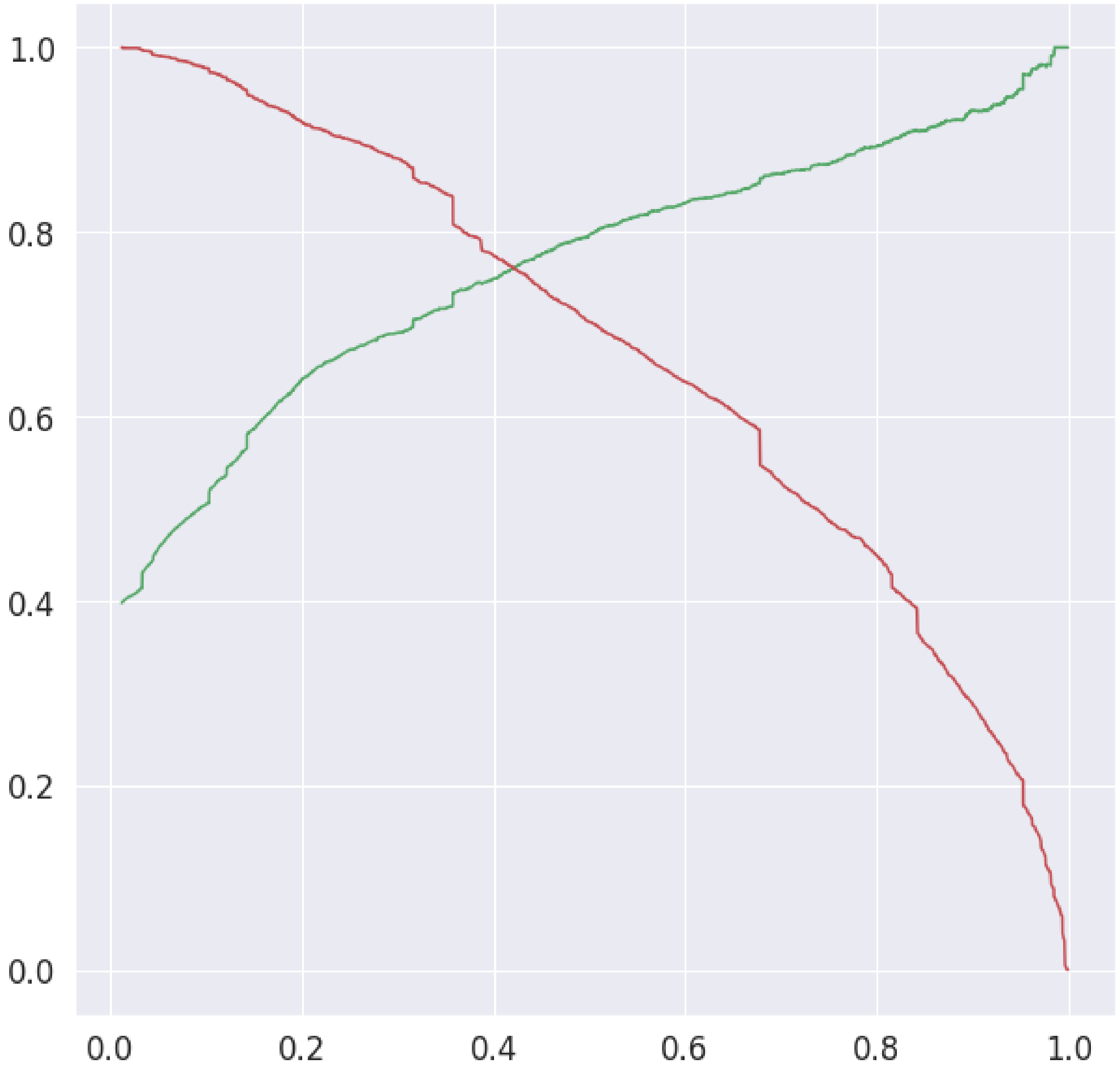

Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.90)

ROC curve (Contd.)

- 0.37 is the optimum point to take it as a cutoff probability



Precision and recall trade-off



Summary

- With the help of the model, we can come to know how did the people started enrolling into online course by Visiting websites, Newspaper article, online advertisement
- It also helps us to identify the category of people enrolling into courses. Whether they are students, working professional, unemployed etc
- What matters to choose a course whether its is for better career prospect, Other and Flexibility & Convenience
- The model shows 81% of accuracy, 79% of sensitivity and 82% of specificity
- The true positive rate for the online course have increased upto 90%
- While analysing both Sensitivity-Specificity and Precision/Recall metrics, the cut-off is based on Sensitivity-Specificity for calculating the final predictions
- As per our understanding, the above model is quite good