LEAD SCORING CASE STUDY

Group Members

1. Gunjit Kapoor
2. Amrita Roy
3. Bidhan Chandra Roy

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

BUSINESS OBJECTIVE

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

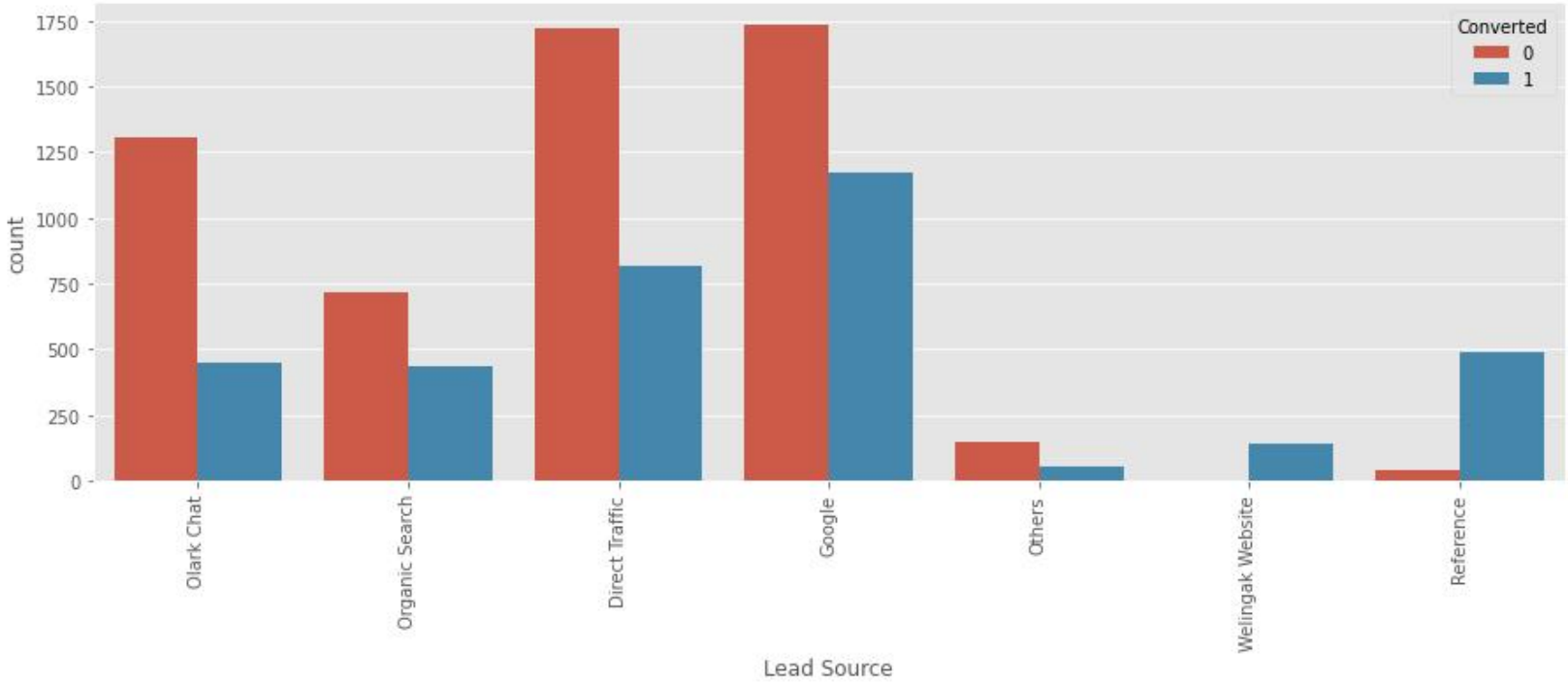
Solution Methodology

- 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.
- EDA
- 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.

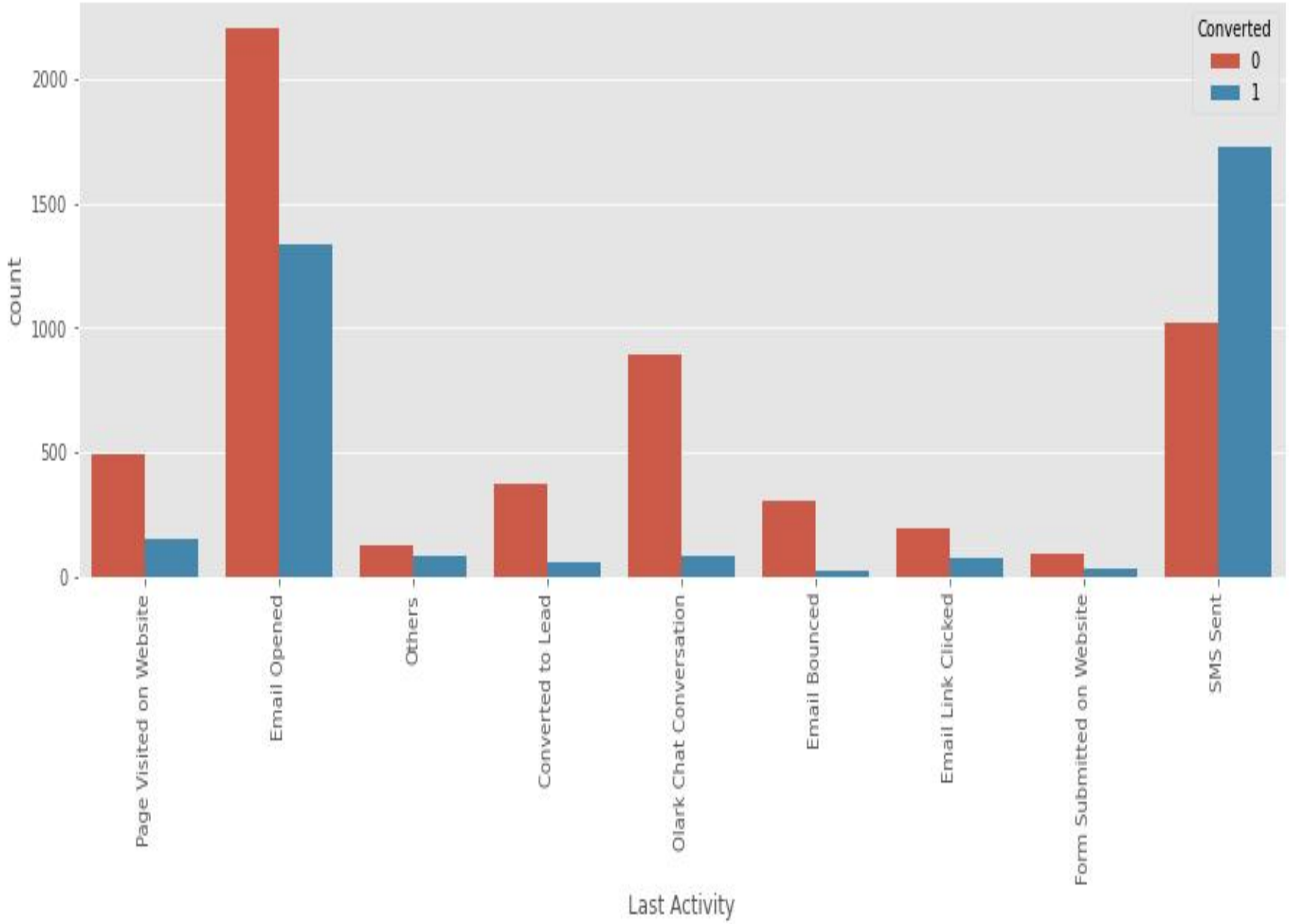
Data Manipulation

- 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 40% as missing value such as ‘How did you hear about ‘X Education’ ,‘Lead Profile’ and ’Lead Quality’.

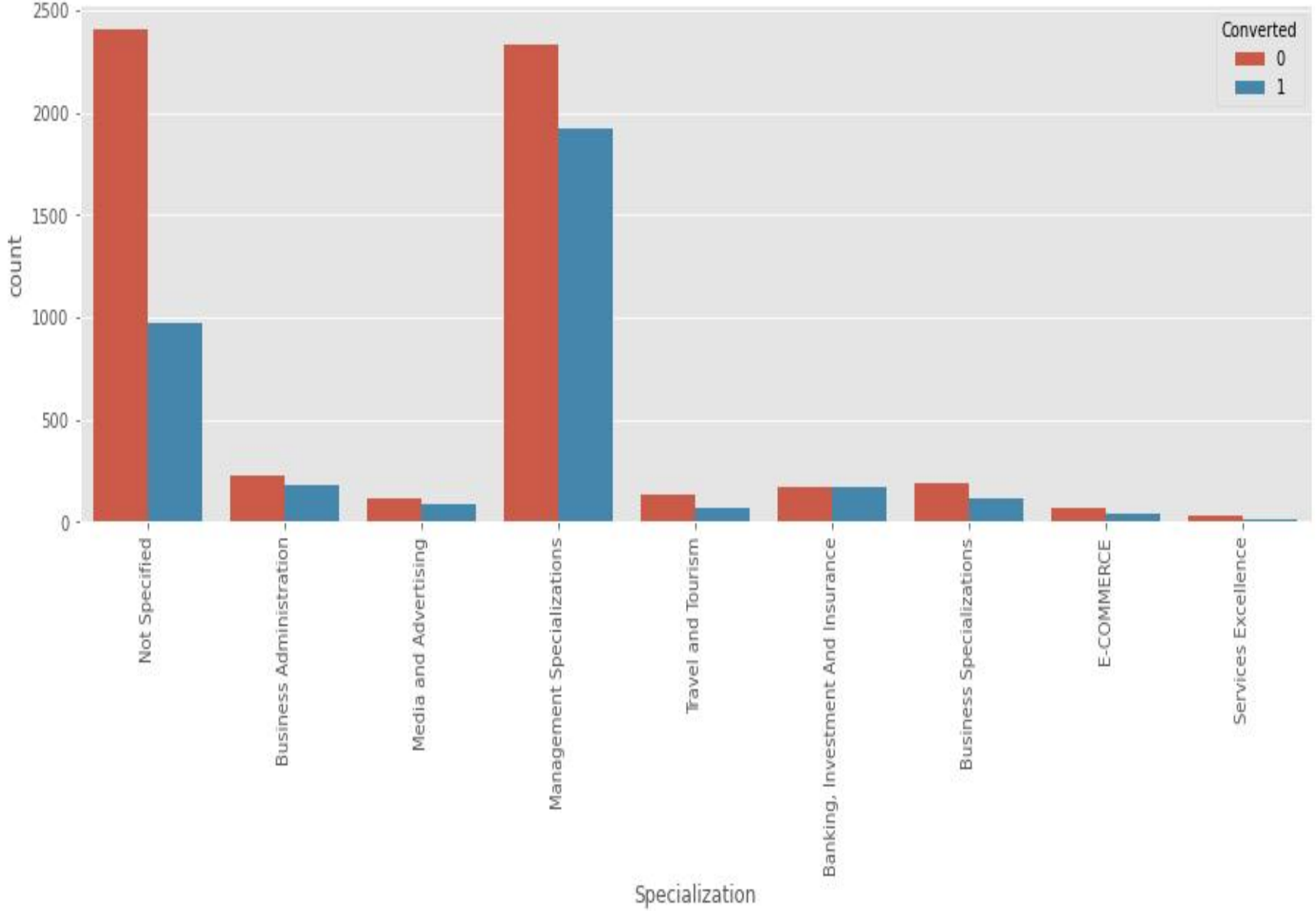
EDA


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Others
Welingak Website
Reference
Lead Source

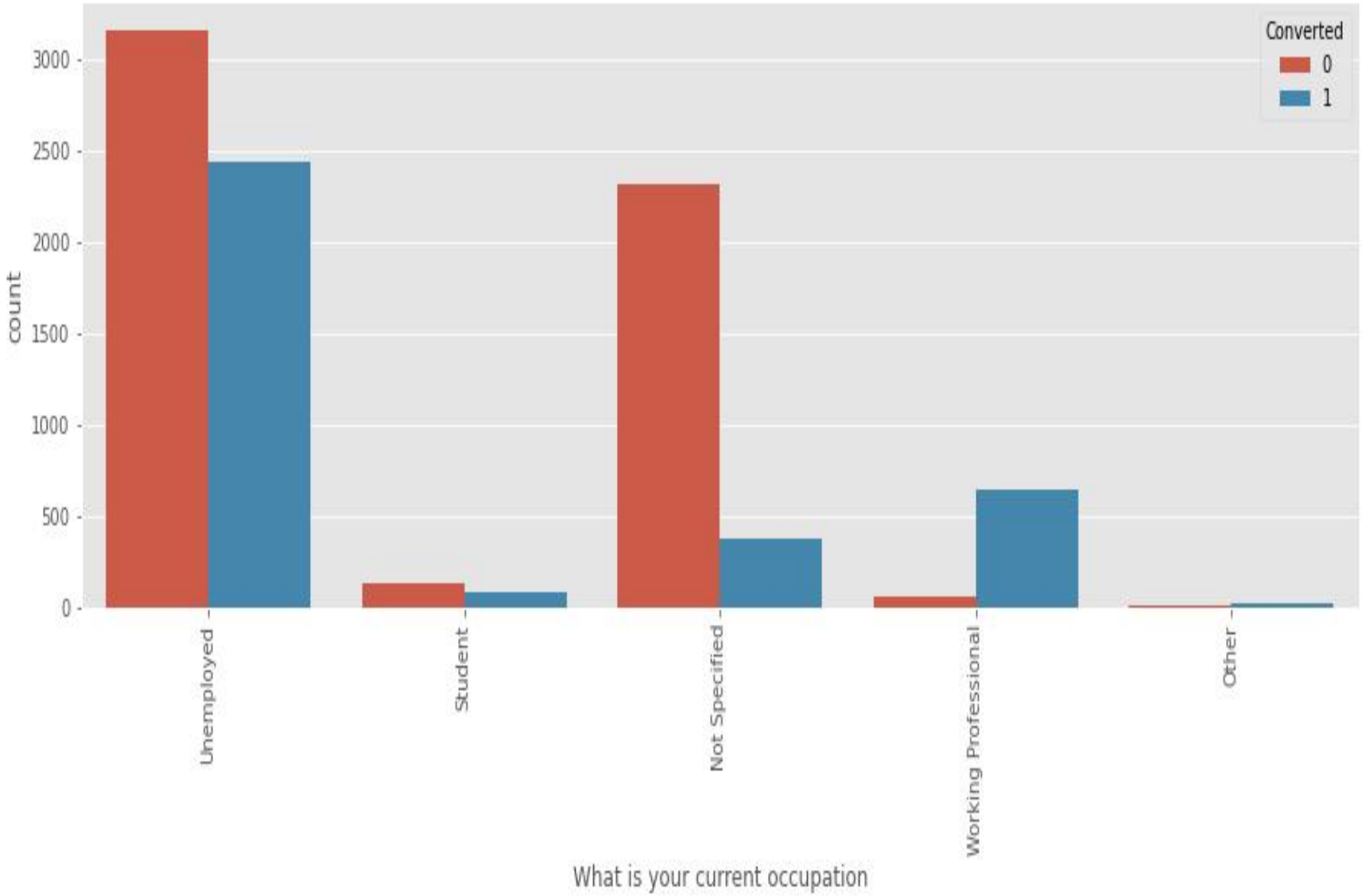
- From the above plot and Lead Source conversion summary, we can infer that:
- Google and direct traffic generates maximum number of leads but has conversion rate of 40% and 32%.
- Welingak website and References has highest conversion rates around 98% and 93% but generates less number of leads.
- olark chat and organic search generates significant number of leads but their conversion rate is around 26% and 38%.¶
- Lead source in 'others' category generate very less leads.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic and google lead source .Also , generate more leads from reference and welingak website since they have a very good conversion rate


Converted
0
1
2000
1500
1000
500
0
count
Page Visited on Website
Email Opened
Others
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
SMS Sent
Last Activity

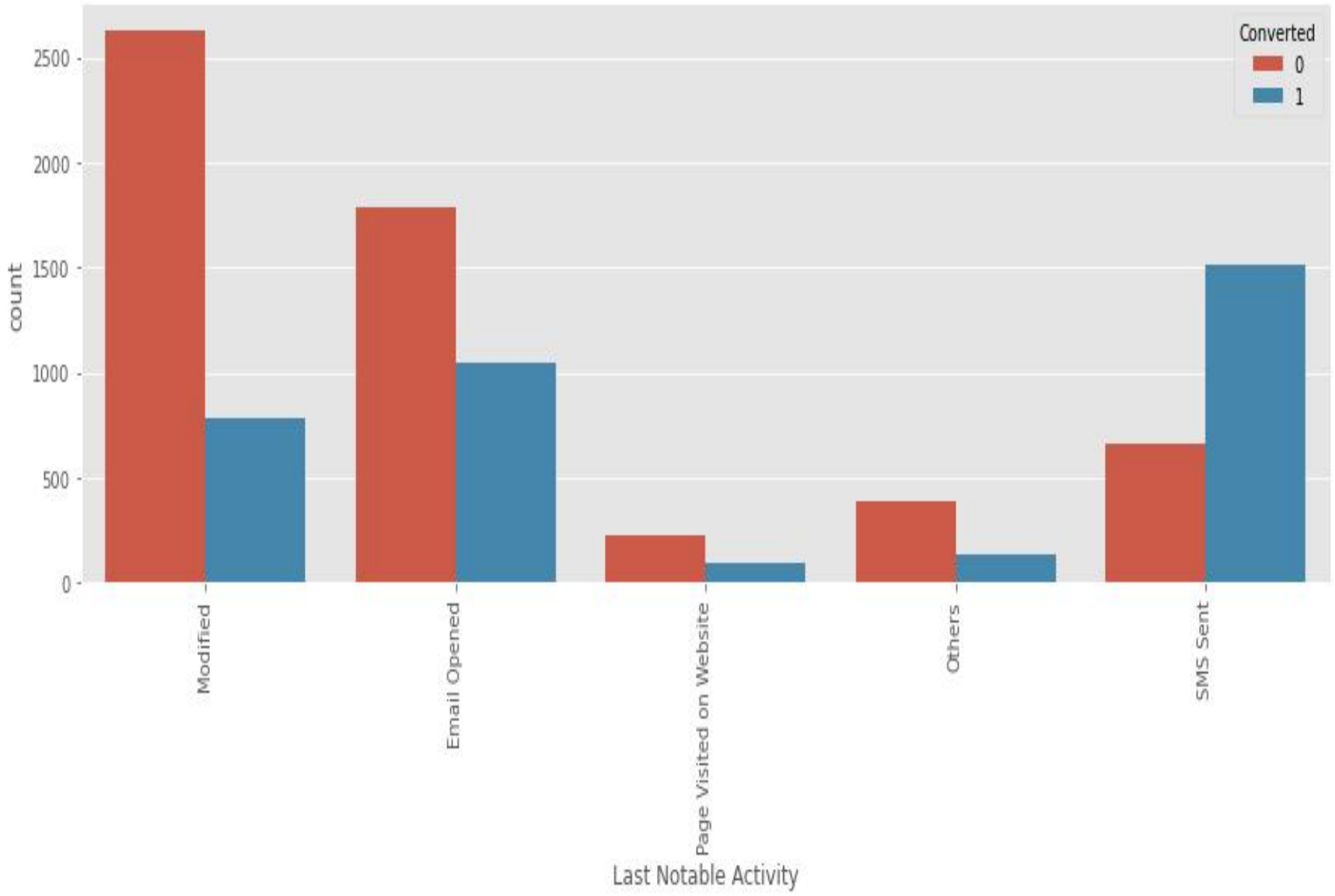
- Maximum leads are generated from people with last activity - SMS sent,Email Opened and others('Visited Booth in Tradeshow', 'Resubscribed to emails','Email Marked Spam', 'Email Received', 'View in browser link Clicked', 'Approached upfront','Had a Phone Conversation', 'Unsubscribed', 'Unreachable')¶
- Conversion rate is around 63%, 38%,41%.
- Conversion rate in case of Email link clicked is 27 % while through 'Form submitted on website' and 'Page visited on website' is 24 %.
- Conversion rate in case of 'Olark Chat Conversation' and 'Email Bounced' is lowest 9% and 8%.


Converted
0
1
count
2500
2000
1500
1000
500
0
Not Specified
Business Administration
Media and Advertising
Management Specializations
Travel and Tourism
Banking, Investment And Insurance
Business Specializations
E-COMMERCE
Services Excellence
Specialization

- **Highest number of lead conversion are from 'Banking,Investment And Insurance','Management Specializations'.**
- **Least conversions are from 'Services Excellence'.**


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Not Specified
Working Professional
Other
What is your current occupation

- **Lead conversion of 'Working Professional' is 92 % which is the highest.**


Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Others
SMS Sent
Last Notable Activity

- **Highest conversion is for 'SMS Sent' which is 69%.¶**
- **The next is 'Email Opened' which has 37%.**

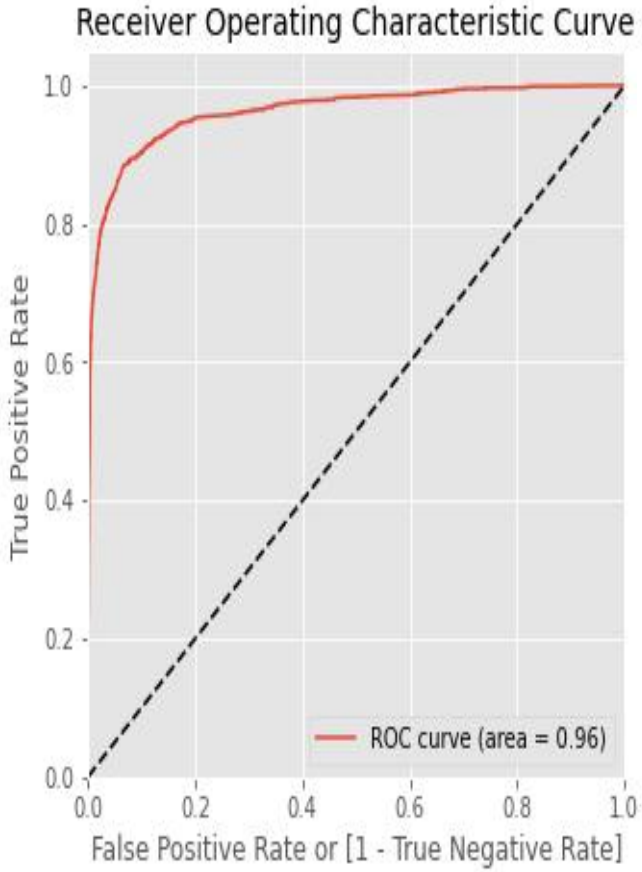
Data Conversion

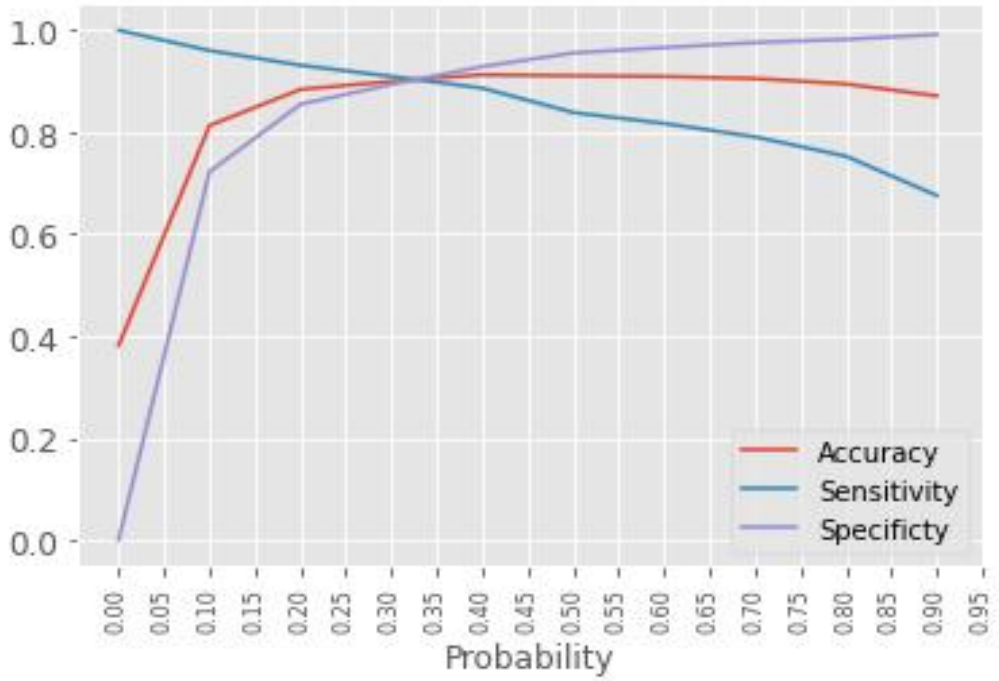
- Numerical Variables are normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9090
- Total Columns for Analysis: 48

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 20 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 90%

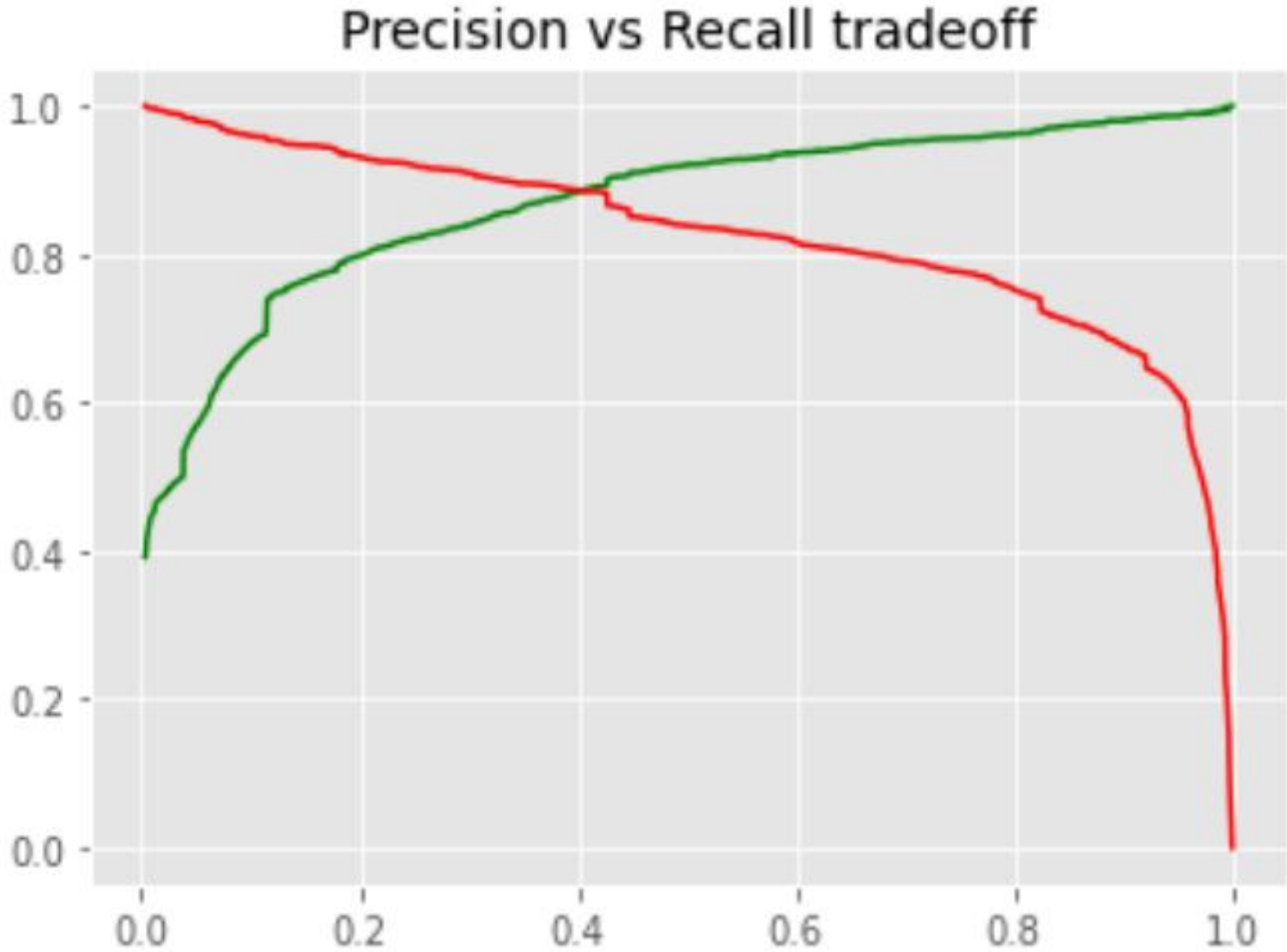
ROC Curve


Receiver Operating Characteristic Curve
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.96)
0.0
0.2
0.4
0.6
0.8
1.0


Accuracy
Sensitivity
Specificty
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95

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 0.325.
- ROC area is 0.96 which is close to 1 which is good indication.


Precision vs Recall tradeoff
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

- It was found that the variables that matter the most in the potential buyers are (In descending order):
- 1. The Tags:
- a. Closed by Horizzon
- b. Will revert after reading the email
- c. Not Specified
- 2. Lead Source
- a. Welingak Website
- b. Olark Chat
- 3. Lead Origin_Lead Add Form
- 4. Last Notable Activity_SMS Sent
- 5. What is the current occupation
- a. Working Professional
- b. Student
- c. Unemployed
- 6. Total Time Spent on Website
- 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.