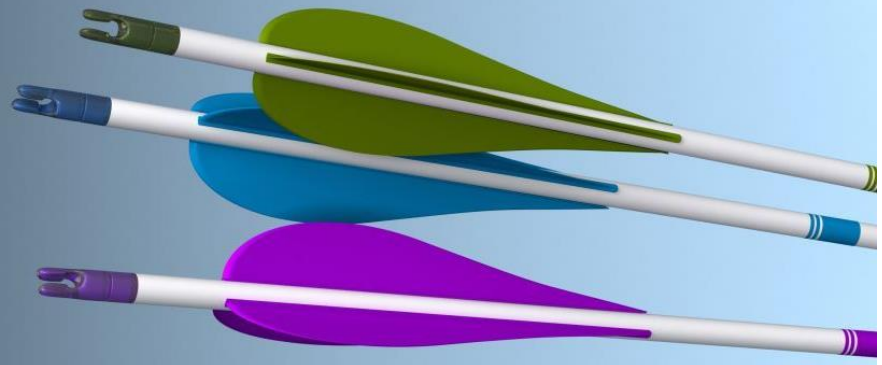
Lead score case study

Group Members:

1. Dharshan Jain.
2. Madaparthi Dileep Sai.
3. Harshal Bhandari.



Problem statement

- An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business objective

- Construct a logistic regression model to allocate lead scores ranging from 0 to 100, intended for the company's targeted approach to potential leads.
- A higher score signifies a "hot" lead, indicating a higher probability of conversion, while a lower score suggests a "cold" lead with a lower likelihood of conversion.
- This scoring system aids the company in prioritizing and tailoring efforts towards leads with a higher potential for conversion.

Solution Methodology

Data cleaning and data manipulation.

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

EDA

- Univariate data analysis: value count, distribution of variable etc.
- 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.

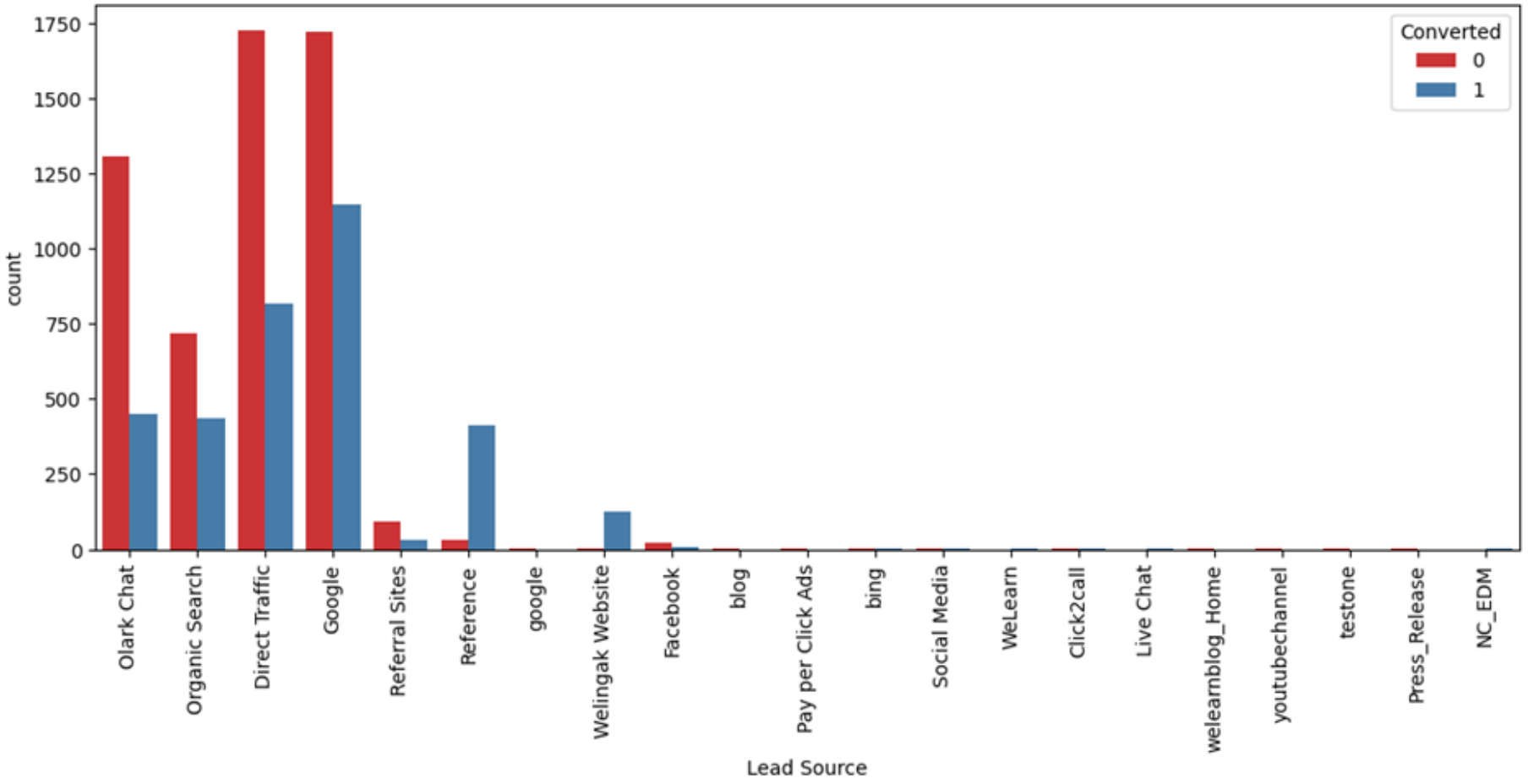
Model presentation.

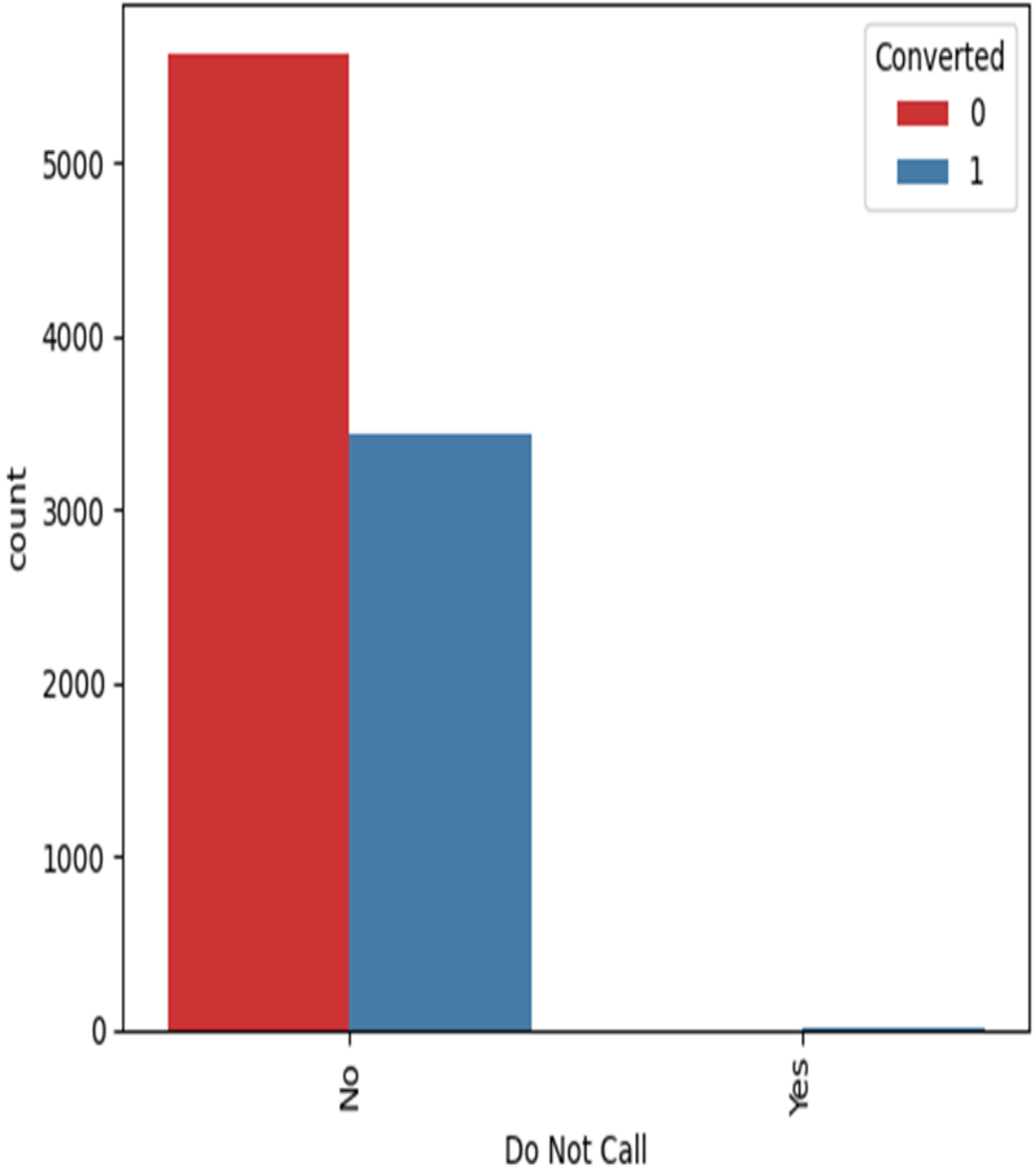
Conclusions and recommendations.

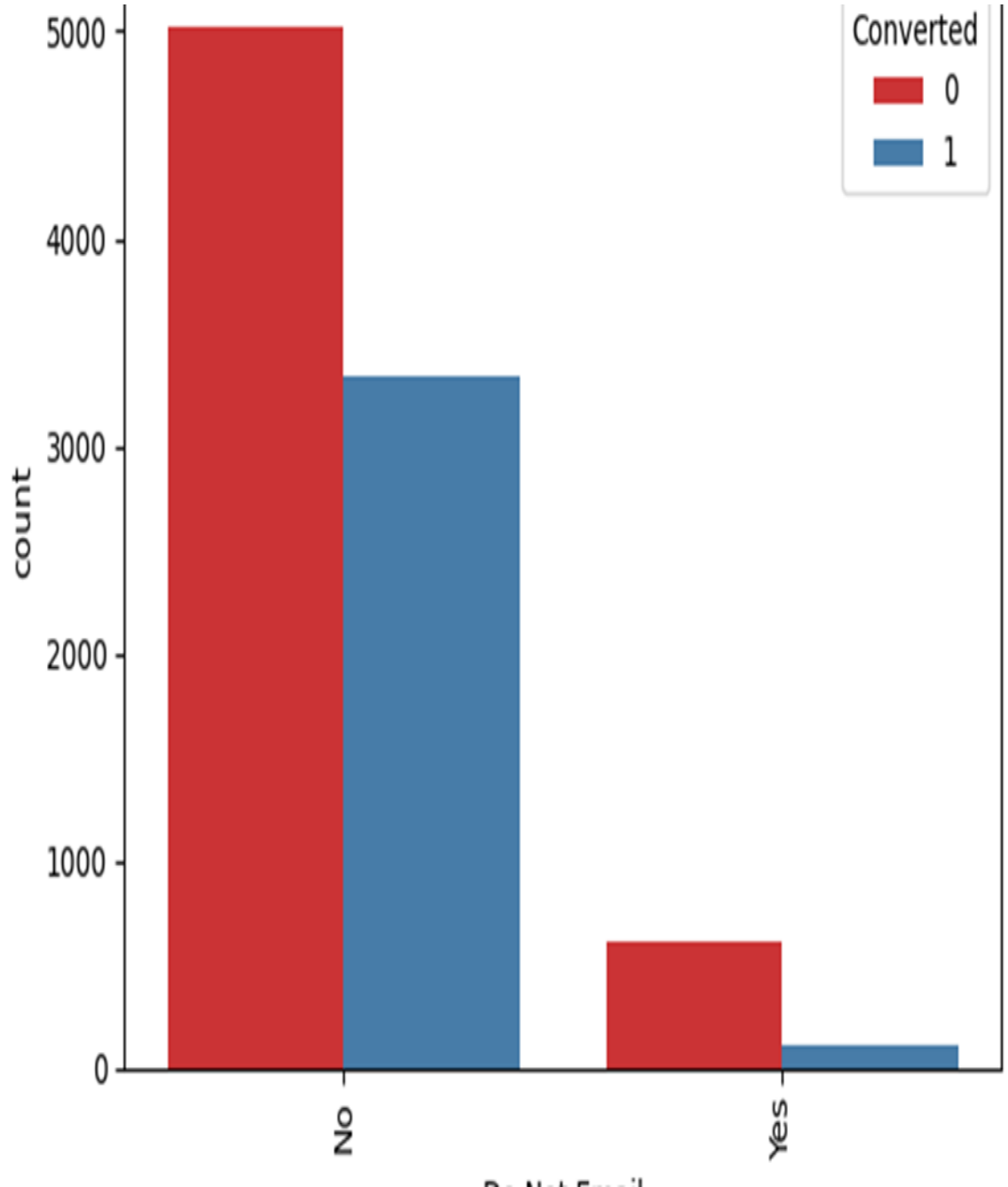
Data Manipulation

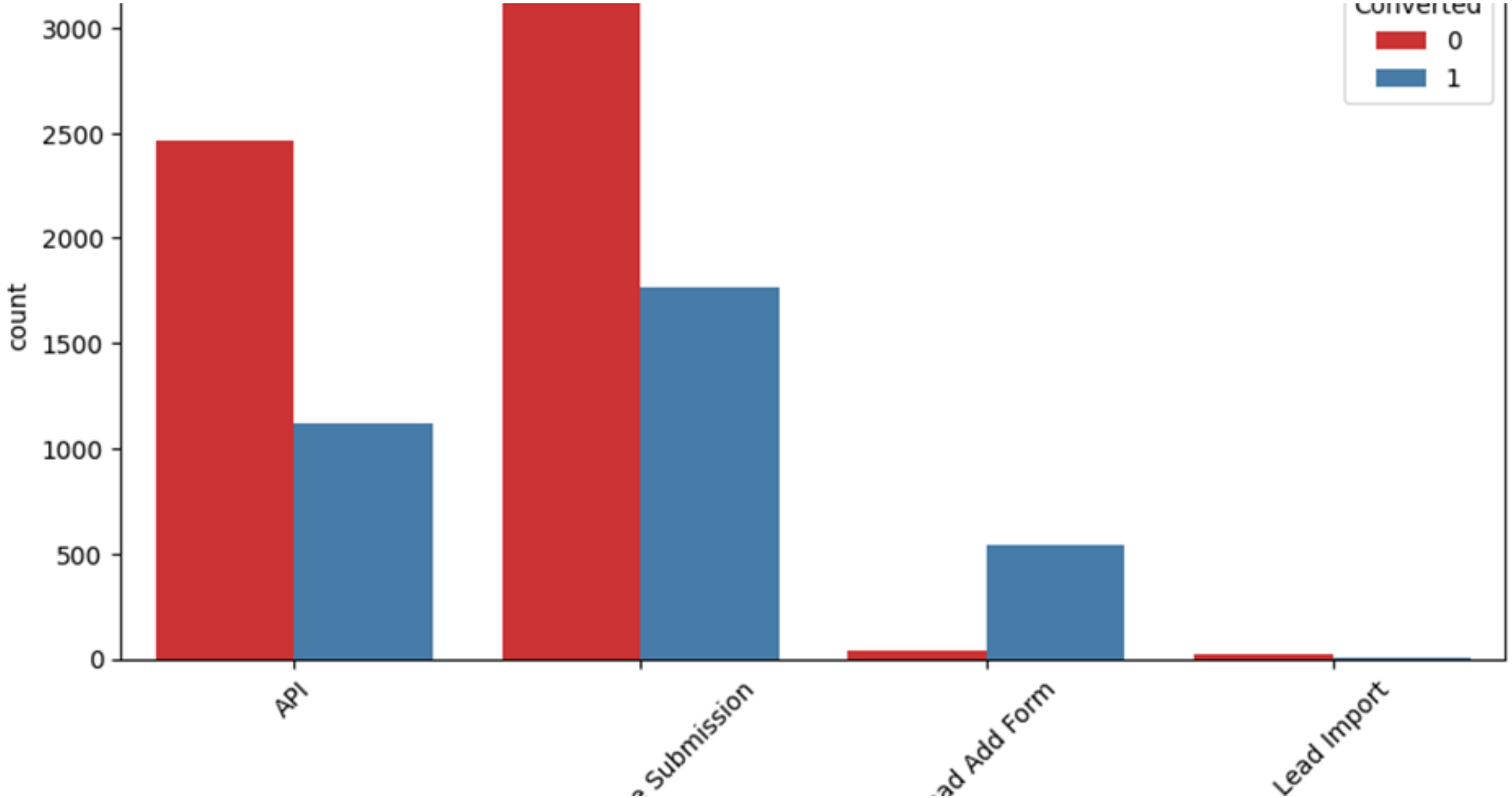
- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect 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’.

EDA


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Call


Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Lead Import

Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

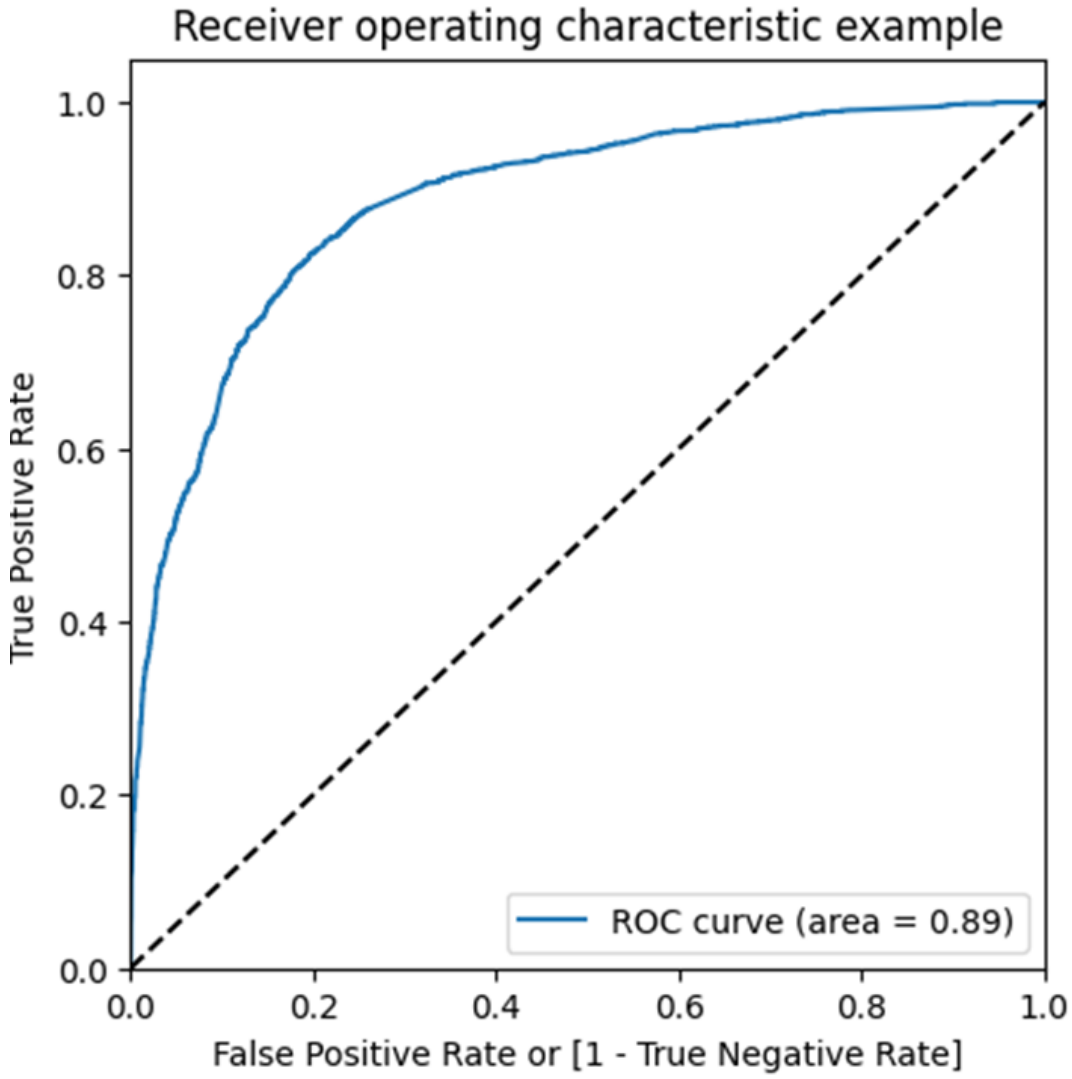
Model Building

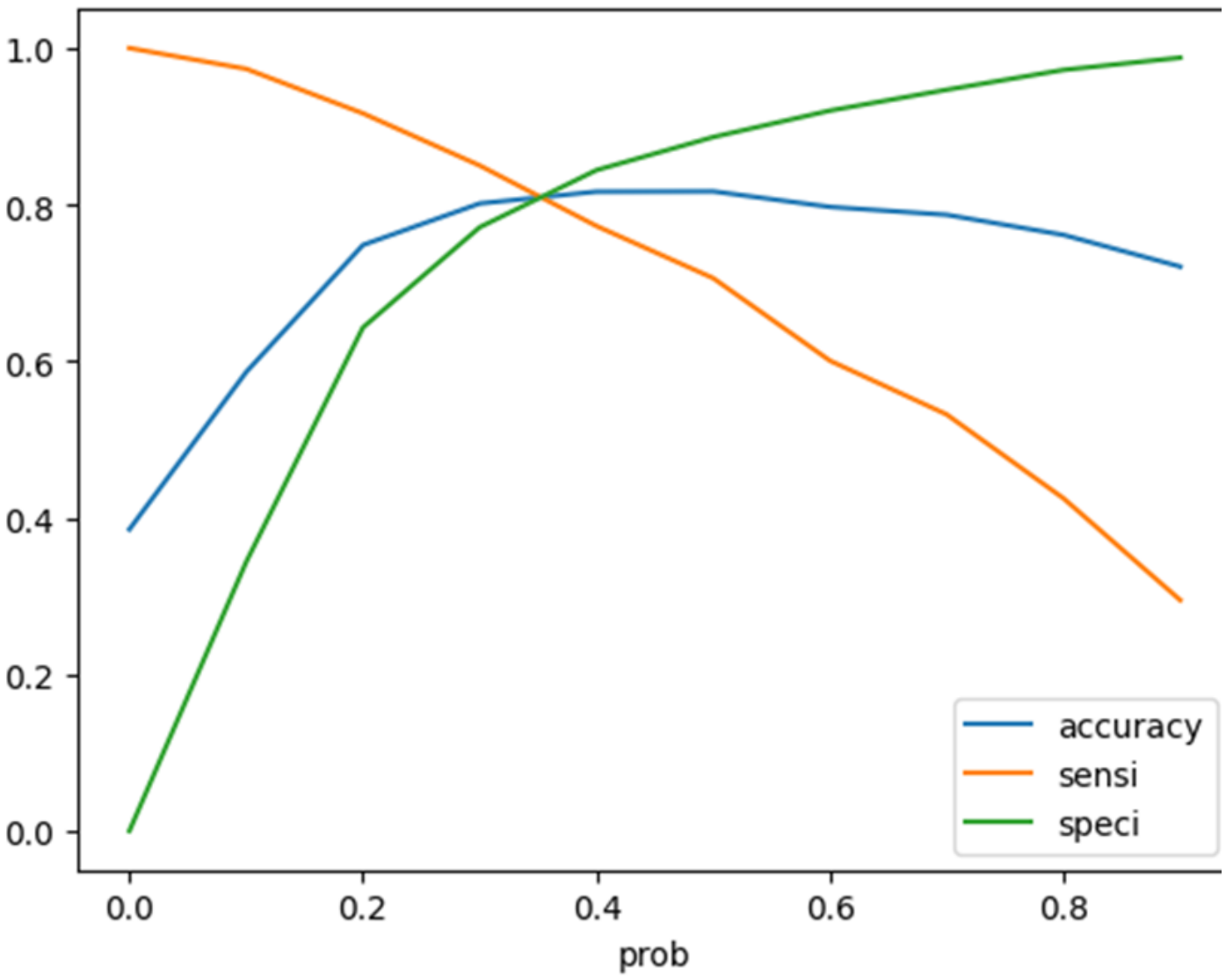
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05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





Optimal Cut off Point

0.34 is the optimum point to take it as a cutoff probability.

Conclusion

Targeted Calling Strategy:

- Prioritize calls to leads from the following sources:
- Welingak Websites
- Reference
- Olark Chat

Focus on leads who:

- Are working professionals
- Have spent more time on the websites
- Showed the last activity as SMS Sent

Avoid calling leads with the following characteristics:

- Last activity recorded as Olark Chat Conversation
- Lead origin being Landing Page Submission
- Specialization listed as Others
- Selected "Do not Email" as "yes"