Lead score case study

BY:

Aditya mehra

Naveen kumar

Nimesh Shroti

Problem Statement

- X education sells online courses to industry professionals.
- X education gets a lot of leads, its lead conversion is very poor though. 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 conversion rate should go up as the sales team will now be focusing more on communication with the potential leads rather than making calls to everyone.

Business objectives

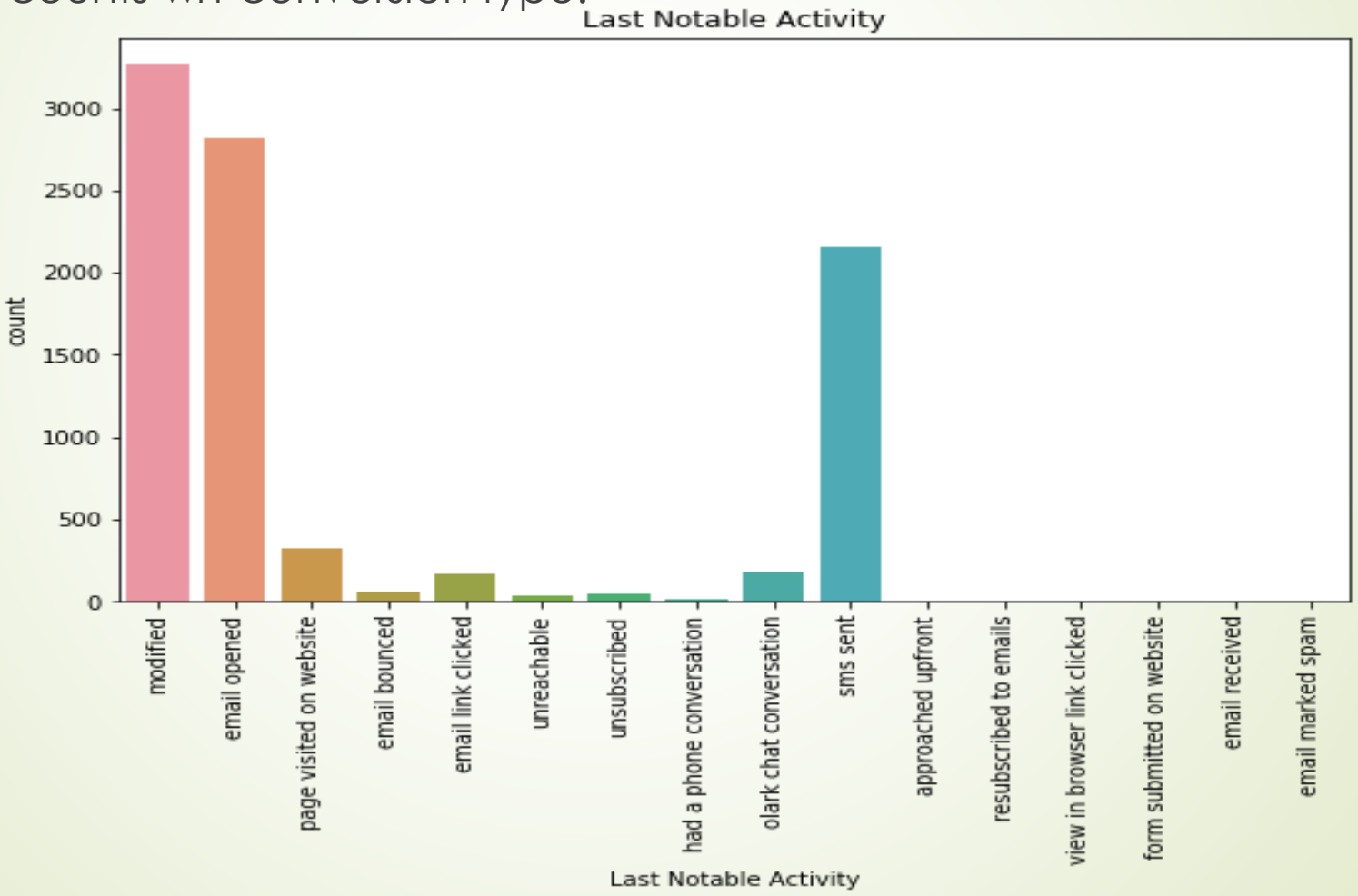
- X education wants us to build a model to give every lead a lead score between 0 – 100. So that they can identify the hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion of 80 %.
- They want the model to be able to handle future constraints as well as like peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

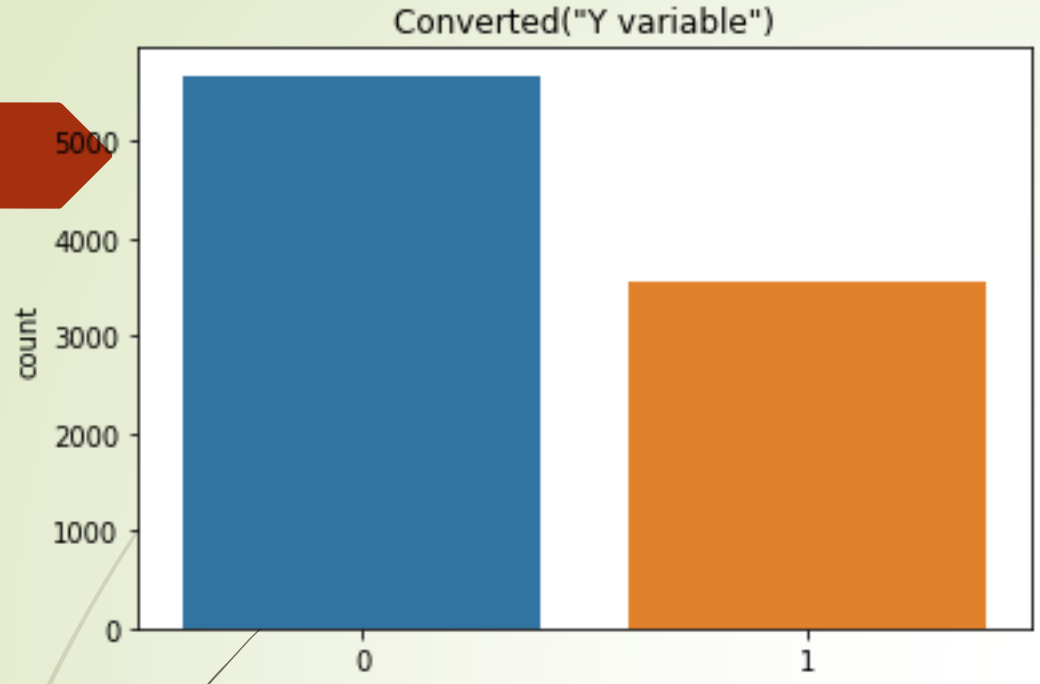
Solution methodology

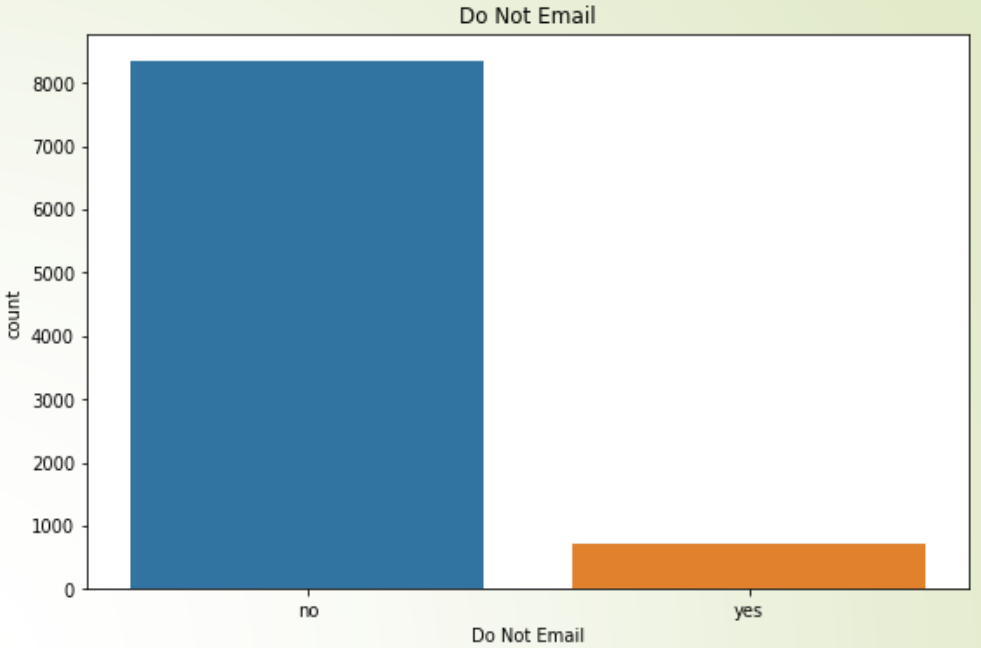
- Data cleaning and data manipulation.
- 1. Check and handle duplicate data.
- 2. Check and handle NA values and missing values.
- 3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
- 4. Imputation of the values, if necessary. 5. Check and handle outliers 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.

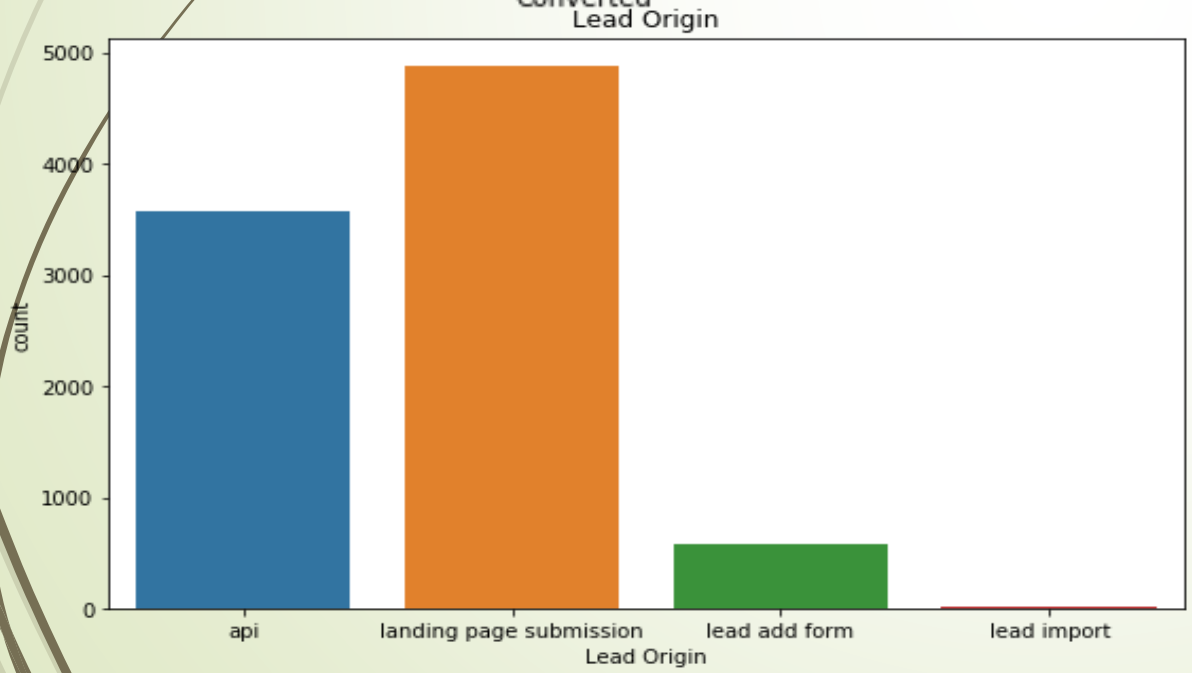
EDA

- Conducted EDA on categorical variables to determine their category wise counts wrt conversion type.

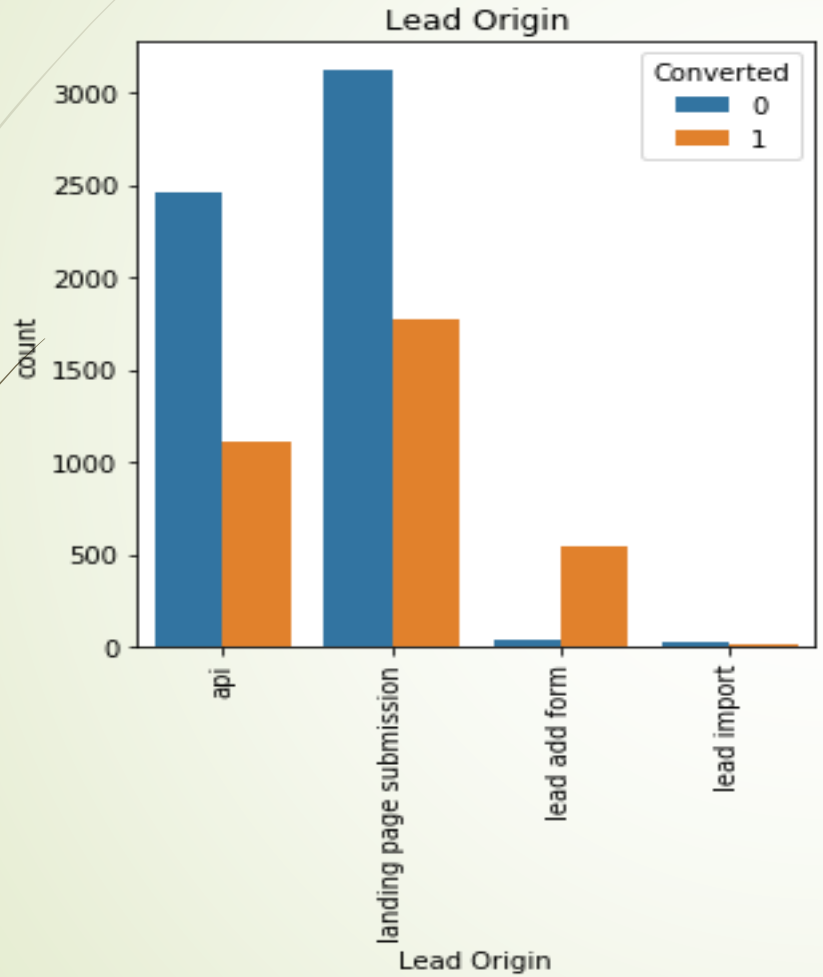


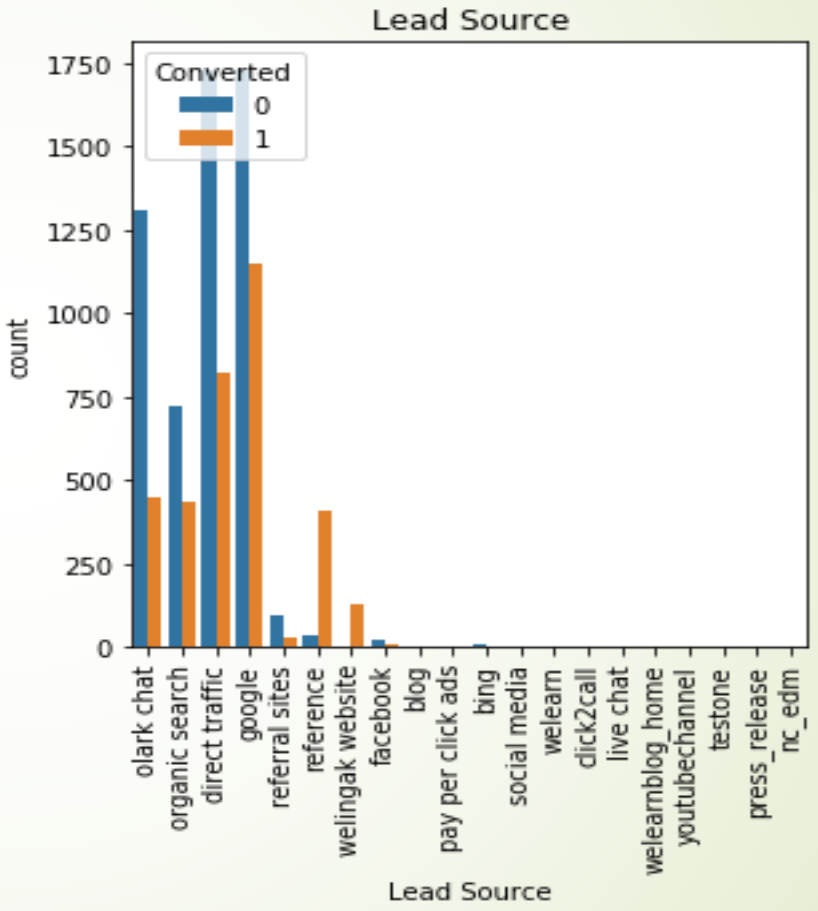

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted


Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email


Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin

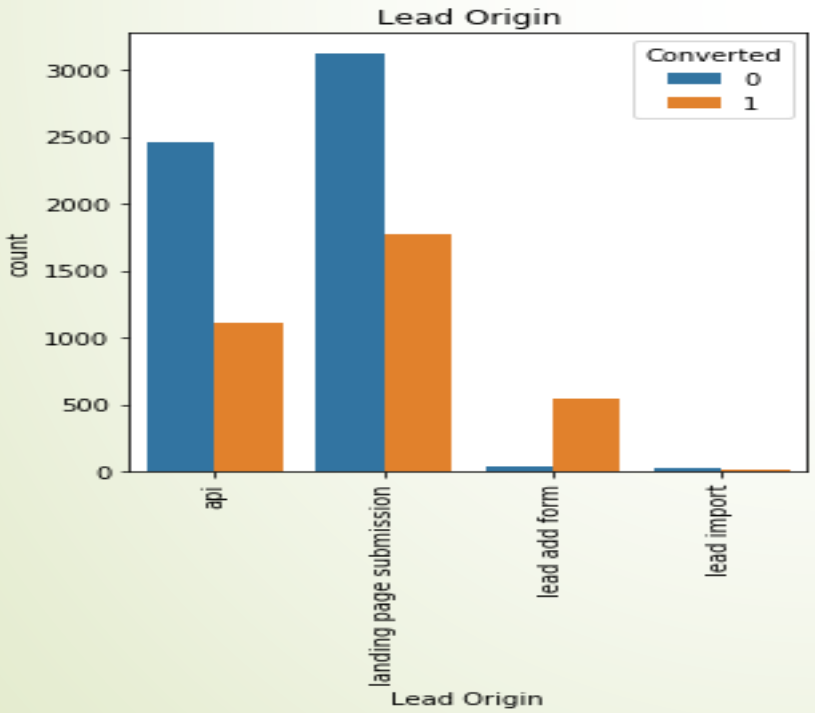
Categorical variable relation

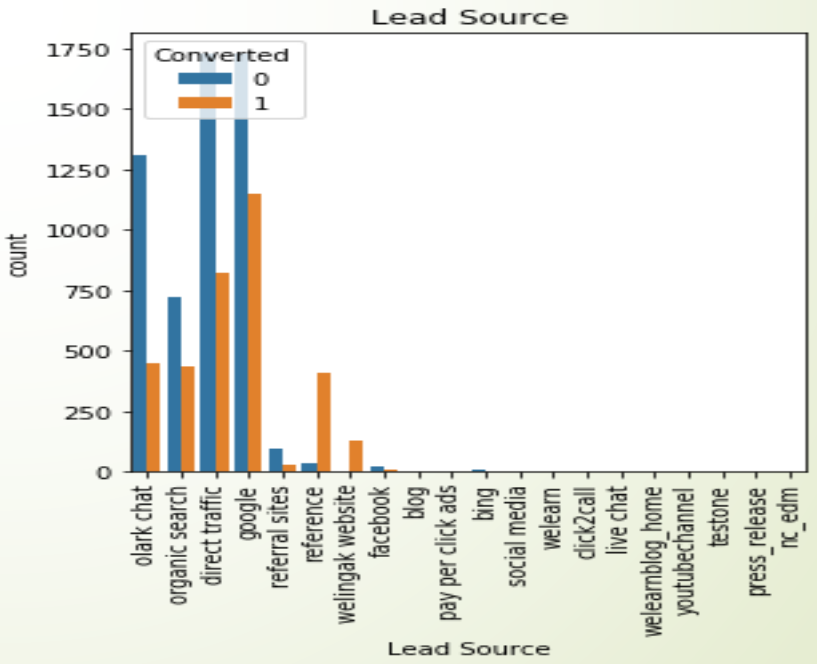


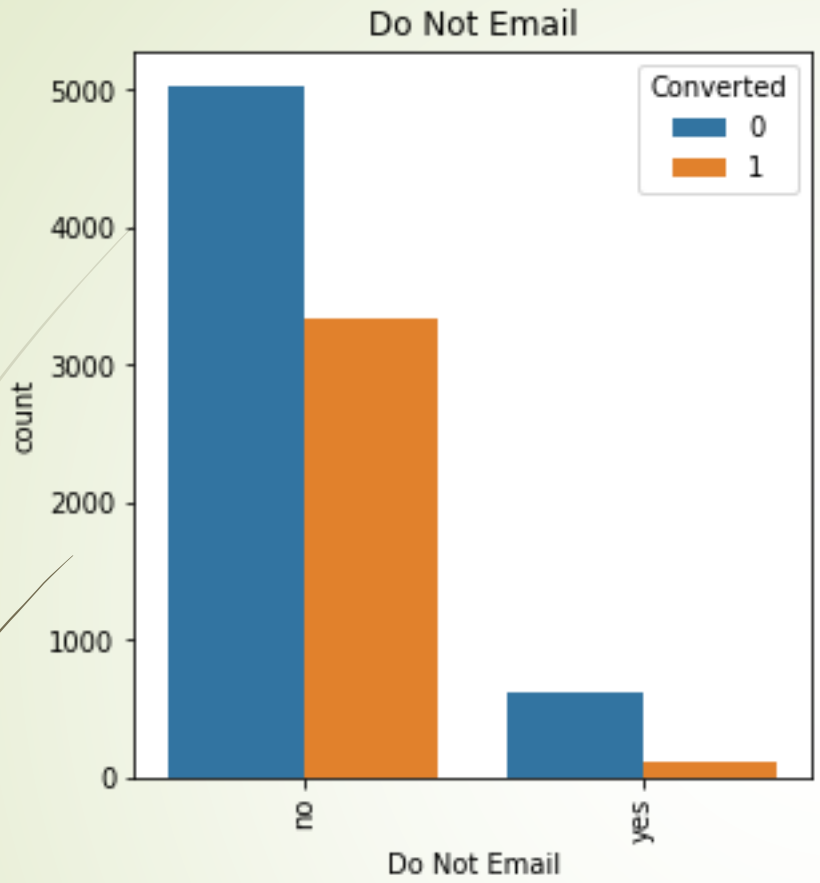


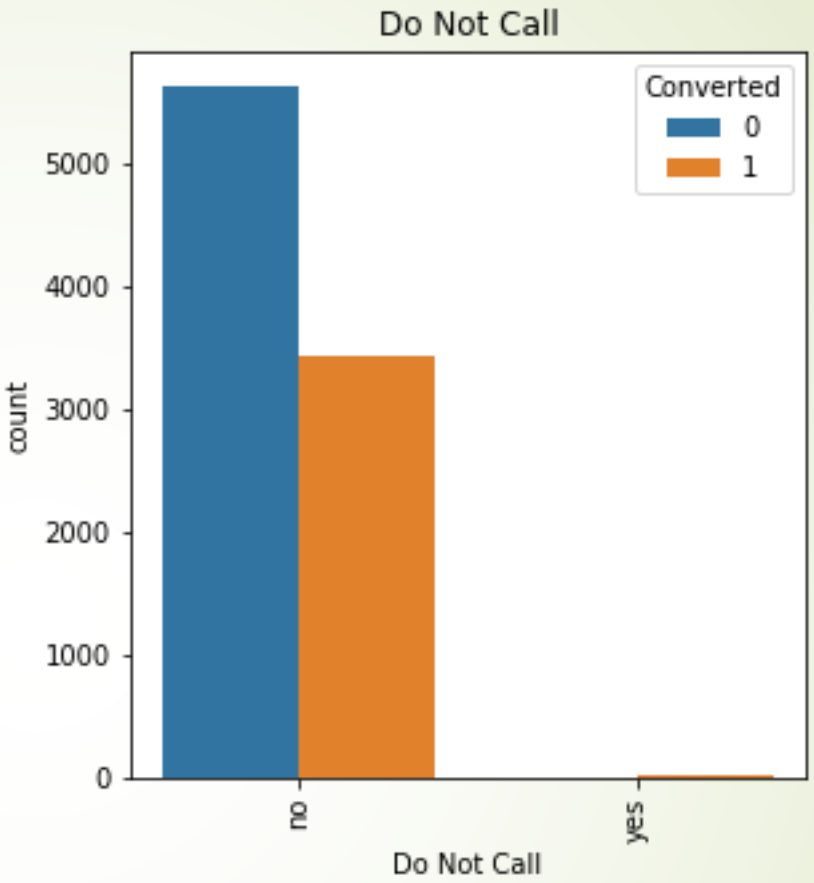
Data conversion

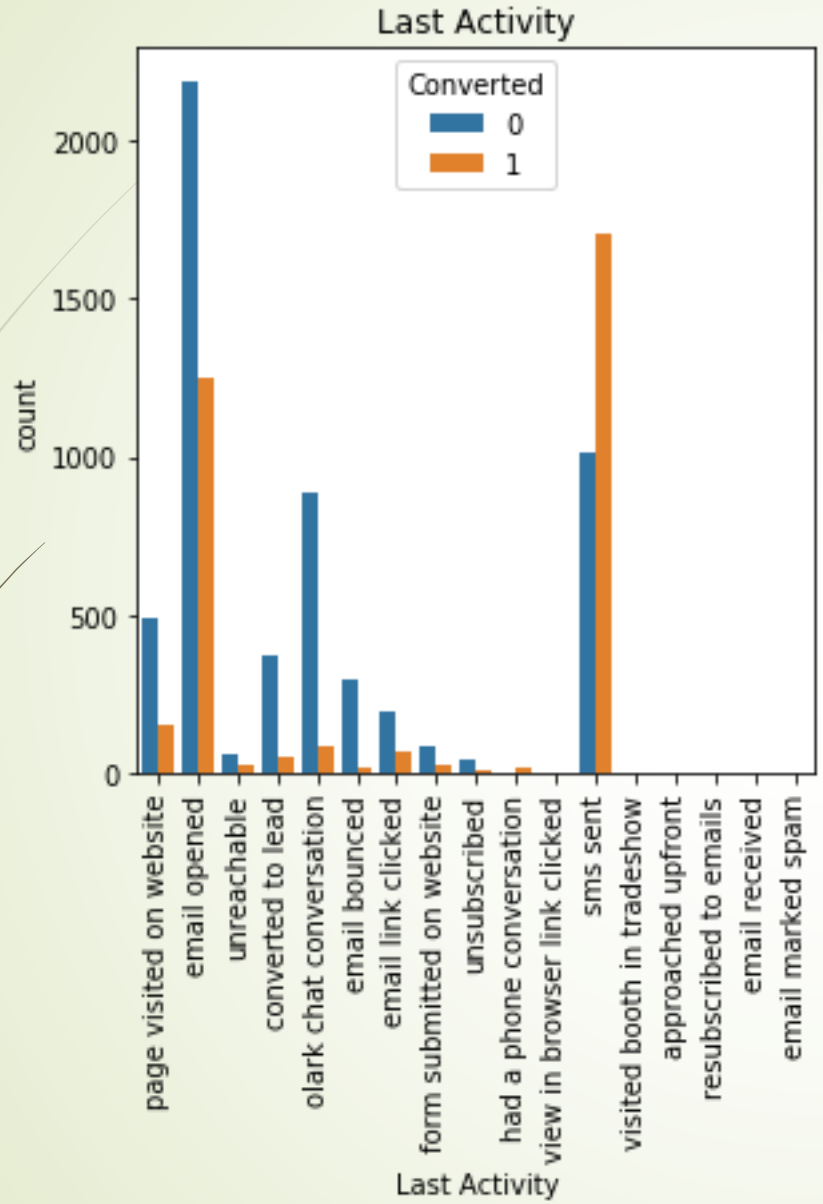
- Generate dummy variables for categorical columns.
- Split the data into training and testing sets with a 70:30 ratio.

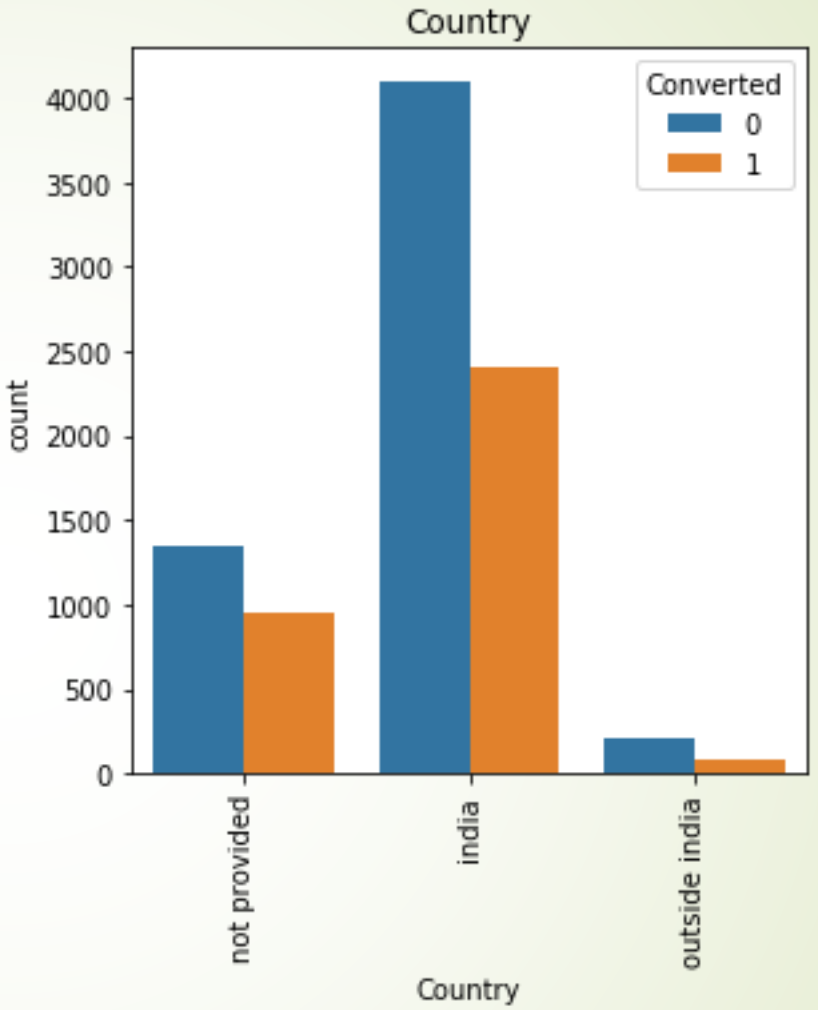





Do Not Email
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Email


Do Not Call
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Do Not Call


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

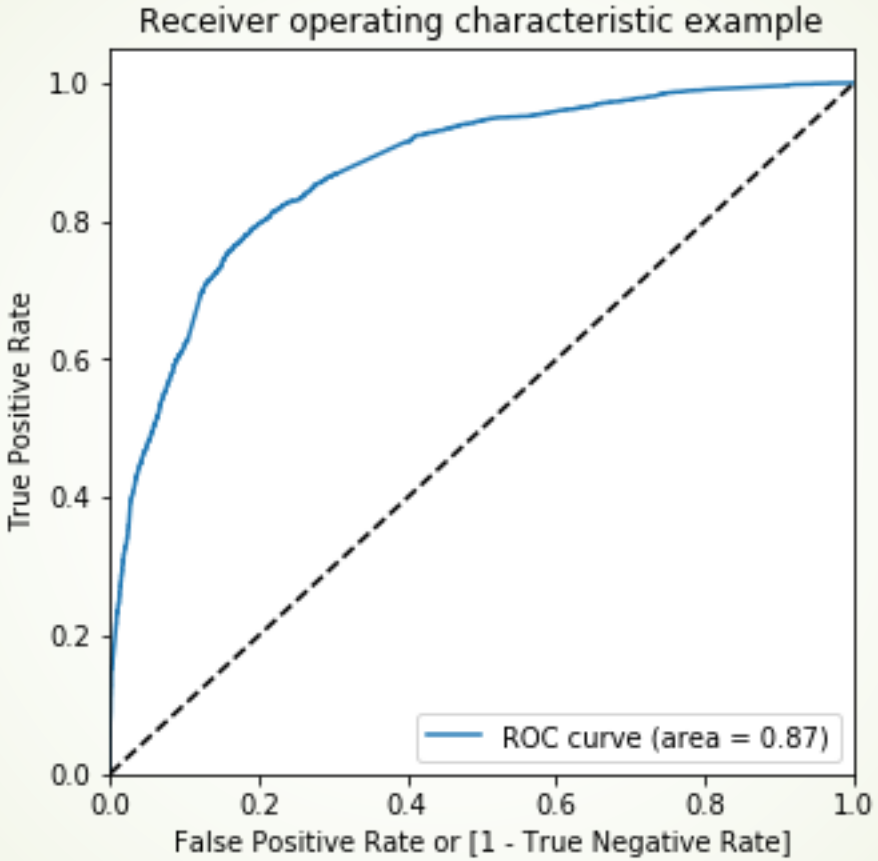

Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

Model building

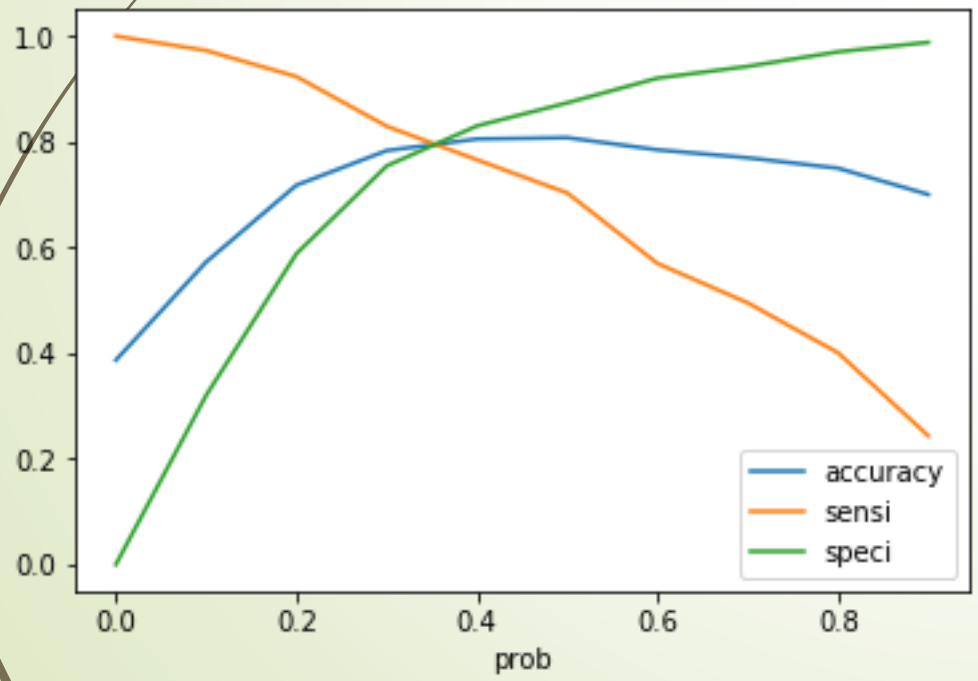
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05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

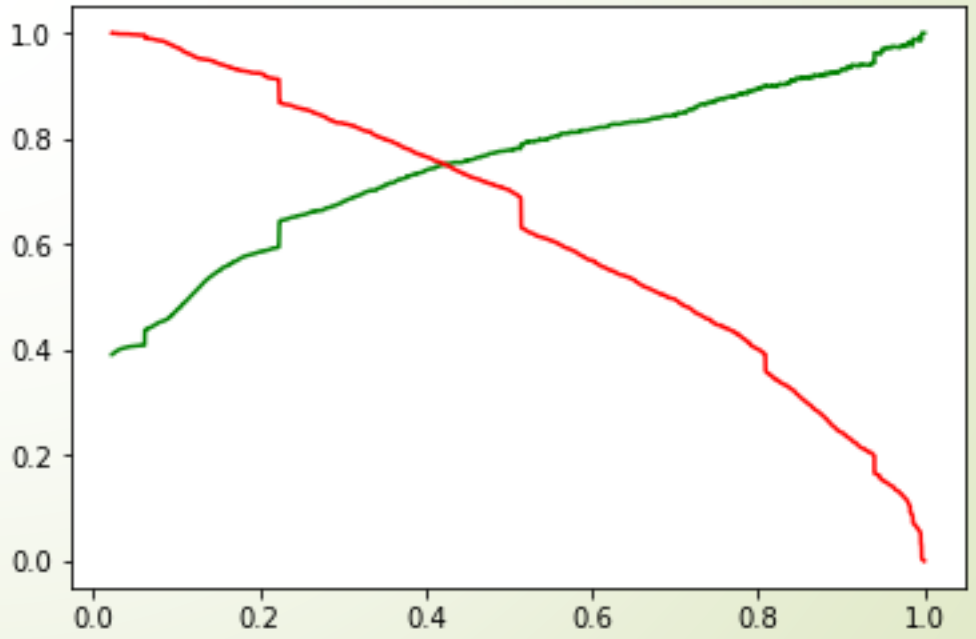
ROC curve

- Generate predictions using the final model on the training dataset.
- Plot the ROC curve for the model, which demonstrates strong performance with the curve closely approaching the top left corner and an AUC of 0.87.
- Use the confusion matrix to compute and plot Accuracy, Sensitivity, and Specificity across different probability thresholds.
- Sensitivity, also known as Recall, measures the proportion of leads that were successfully predicted as converted among those that actually converted.
- Recall = True Positives (TP) / Total Actual Positives (TP + FN)
- Precision refers to the proportion of leads that were correctly predicted as converted out of all the leads predicted as converted.
- Precision = True Positives (TP) / Total Predicted Positives (TP + FP)
- In this scenario, accepting slightly lower precision may lead to contacting some non-potential leads. Therefore, we are willing to sacrifice some precision in favor of achieving higher recall.


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
ROC curve (area = 0.87)

- Our goal is to contact as many leads as possible to maximize the conversion rate. To do this, we need to establish a probability cutoff that prioritizes recall, ensuring that we capture the majority of potential leads (Hot Leads), while still maintaining a reasonable level of precision to avoid excessive calls to non-potential leads.
- Therefore, we set the cutoff probability at 0.35 to strike a balance between higher recall and sufficient precision. Consequently, our model achieved a lead conversion rate of approximately 79%, meaning we successfully contacted 79% of the leads who eventually converted, thereby supporting a strong conversion rate.





Conclusion

- It was found that the variables that mattered the most in the potential buyers are (in desc order) :
  1. Google
  2. Direct traffic
  3. Organic search
  4. Welingak website
- When the last activity was :
  1. SMS
  2. Olark chat conversion
- When lead origin is lead add format.
- When their current occupation is a working professional.