LEAD SCORING CASE STUDY

SUBMITTED BY:
1. YUVRAZ DHUNGANA
2. NEHA MISHRA
3. HARENDER THAKUR

PROBLEM STATEMENT

- Education company
  - X education
  - sells the courses online only to working/industry professionals
- People filling up form gets converted to lead
- Company gets many leads but only 30% get converted
- Aim: Identify the “HOT LEADS”
- Getting HOT LEADS can help increase Sales
- Sales team contact hot leads only

DATA

- Dataset with 9000 data points
- Target Column : “Converted”
- Converted:1
- Not Converted :0
- Categorical Variable : “Select”: as good as null values:

GOALS OF CASE STUDY

1. Building logistic model to assign a score 0 to 100
2. Higher score higher conversion chances
3. Some problems need to be looked at that may affect in future

RESULTS EXPECTED

1. A well-commented Jupyter notebook with at least the logistic regression model, the conversion predictions and evaluation metrics.
2. The word document filled with solutions to all the problems.
3. The overall approach of the analysis in a presentation.
    1. Mention the problem statement and the analysis approach briefly
    2. Explain the results in business terms
    3. Include visualizations and summarize the most important results in the presentation
4. A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

SOLUTION METHODOLOGY

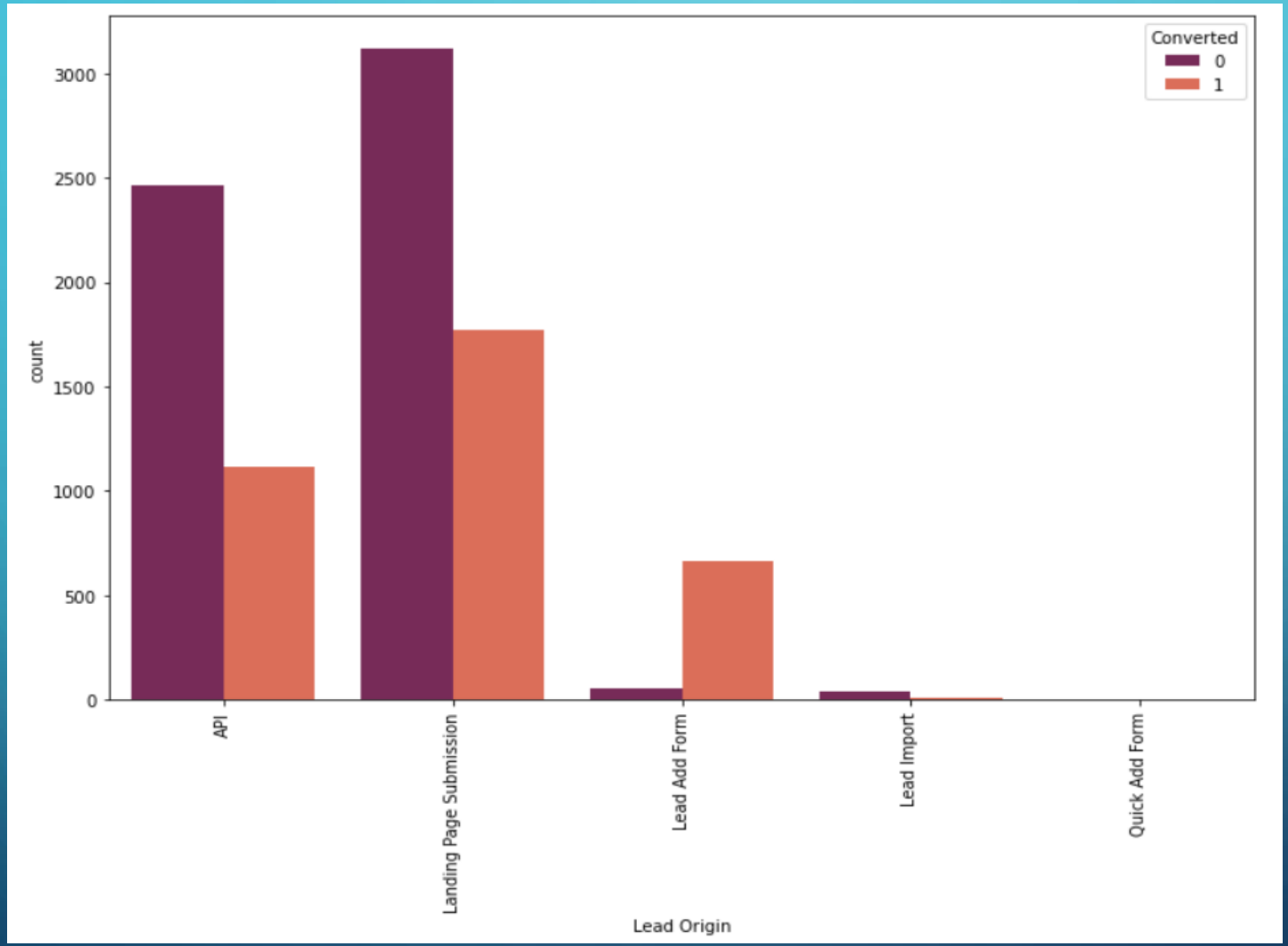
- Data cleaning
  - Finding null values
  - Imputing the null values
  - Dropping the values if it has more than 40% of its total values as null values
  - Checking for outliers
- EDA
  - Univariate Analysis:
  - Bivariate Analysis
- Data Preparation :
  - Created dummy variables using get_dummies() in pandas
  - Created train dataset : 70%;to be used for modelling
  - Created Test dataset :30%; Can be considered as Unseen data to be used for evaluation
  - Feature scaling: mandatory in linear and logistic model

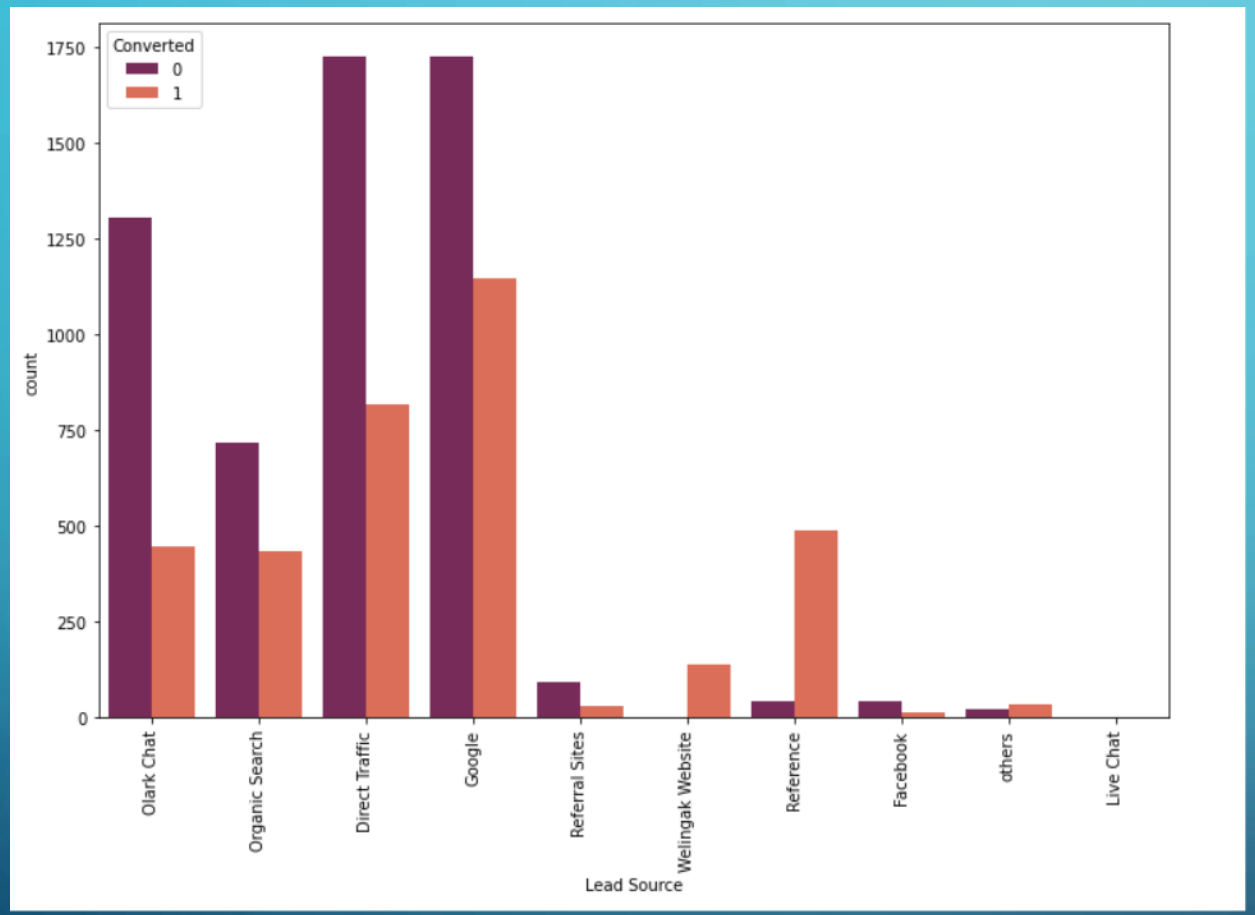
SOLUTION METHODOLOGY CONTINUED

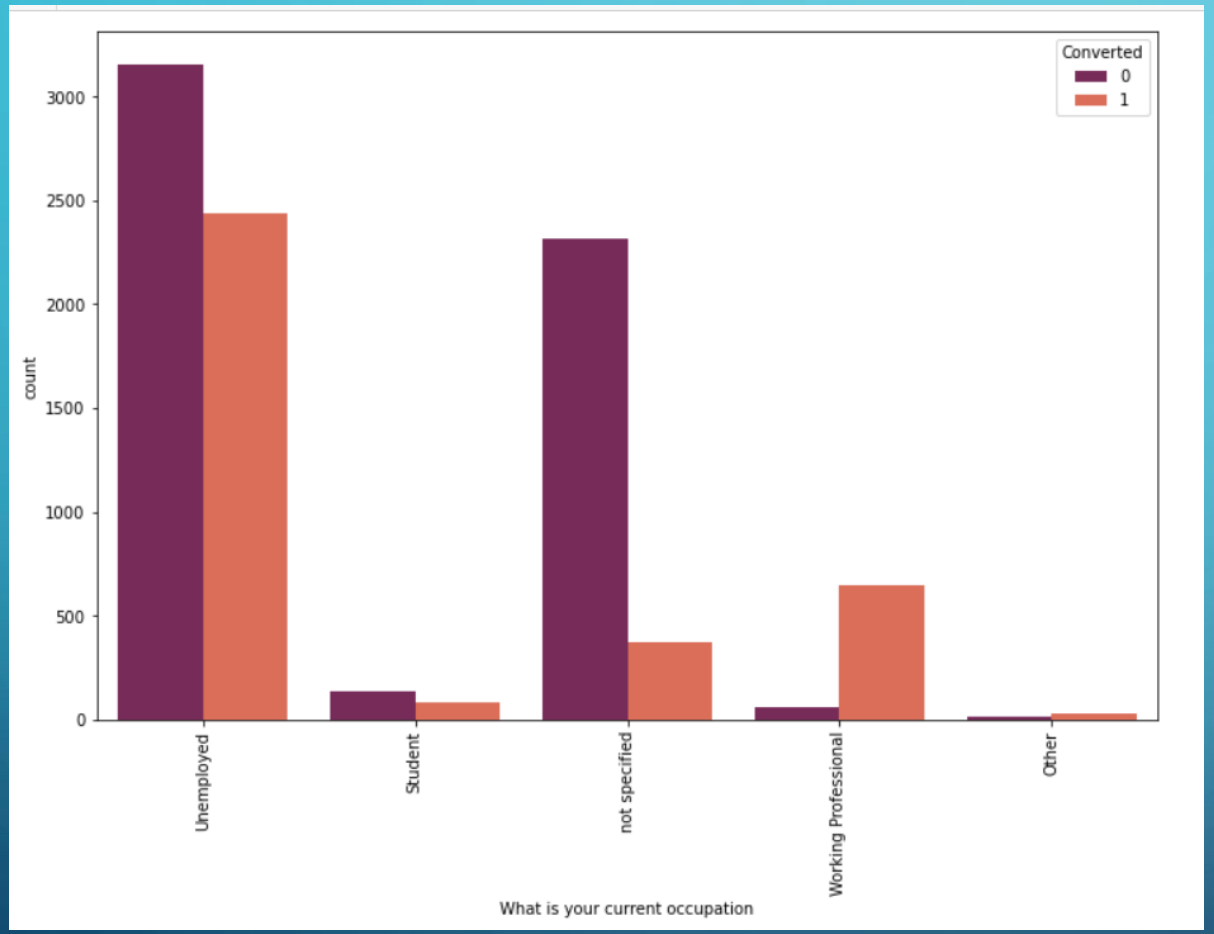
- Classification Technique : Model building executed through Logistic Regression
- Model Validation
- Model Presentation
- Conclusion and Recommendation

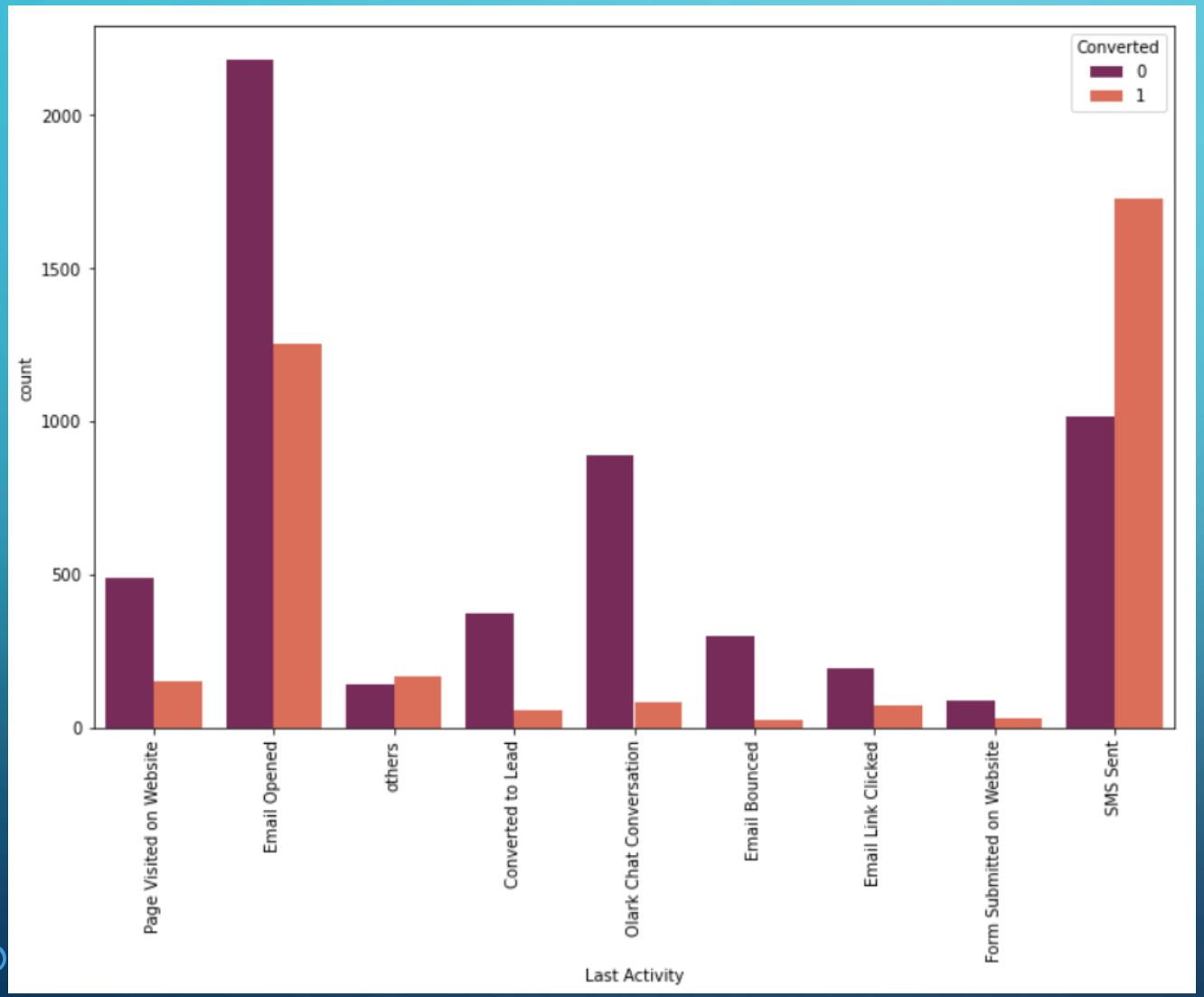
DATA MANIPULATION

- Total rows: 9240, Total columns:37
- Dropped Columns:
  - Single value features:
    - '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', 'Chain content etc
  - 'Prospect Id', 'Lead Numbers' are dropped : Not required for analysis
  - After checking for null values and the columns don't have much variance:
  - 'Do Not Call', 'What matters most to you in choosing the course', 'Search', 'Newspaper', 'Newspaper articles', 'Digital Advertisement', 'X Education Forums', etc
  - Variables with more than 40% null values

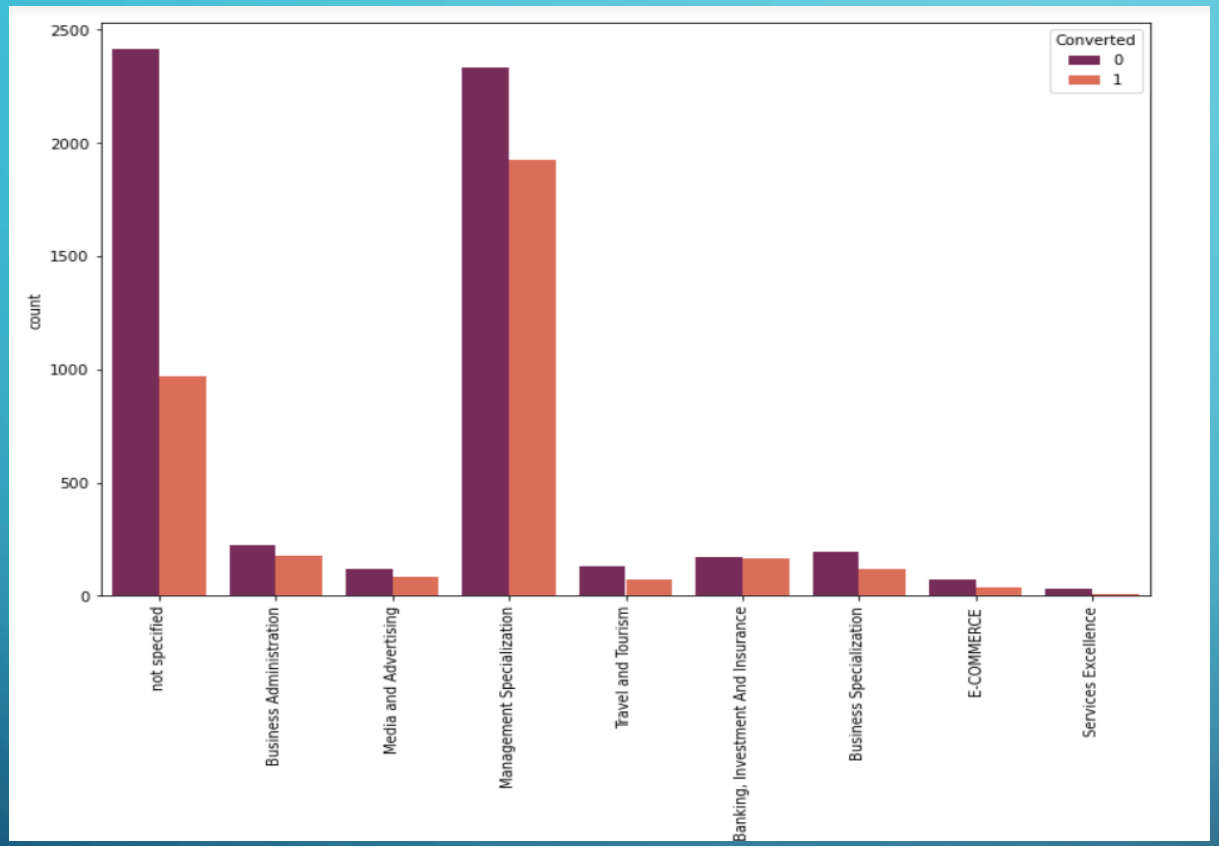

Converted
0
1
3000
2500
2000
1500
1000
500
0
count
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Converted
0
1
1750
1500
1250
1000
750
500
250
0
count
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
others
Live Chat
Lead Source


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


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

SPECIALIZATION GRAPH



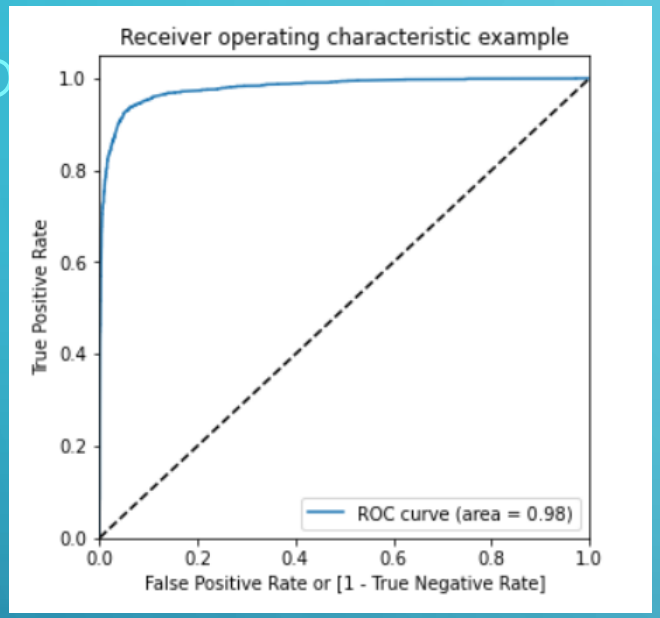
DATA CONVERSION

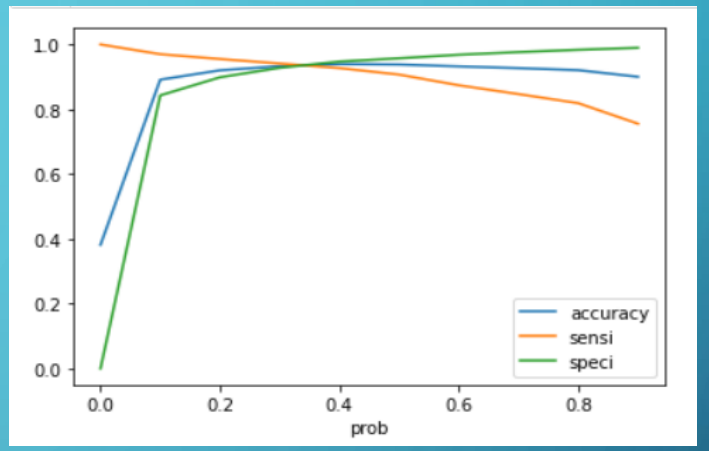
- Numerical values are Normalized
- Dummy variables are created for Object type variable
- Total no. of Rows analysed: 9240
- Total no. of columns analysed: 13

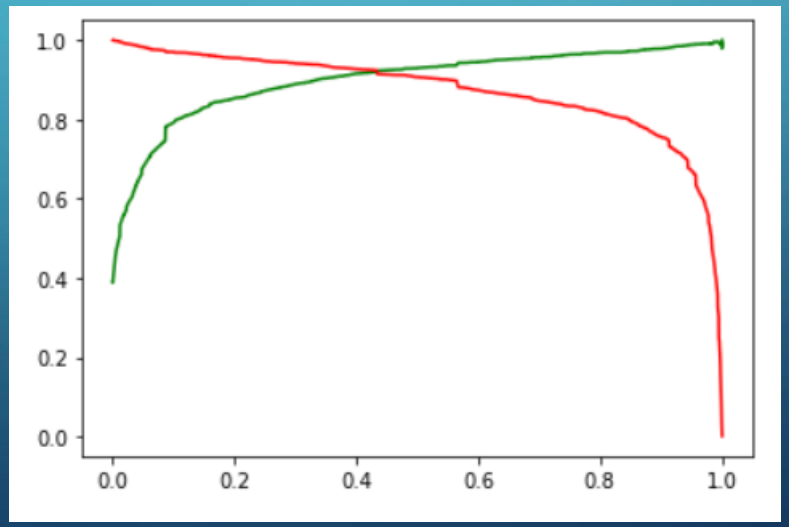
MODEL BUILDING

- RFE technique to perform variable selection(max20)
- Build Logistic Regression model with good sensitivity
- Drop one by one by p-value
- Check p-values and vif
- Find optimal probability cut off
- P-value<0.05 VIF<5
- Recall-> How good the model is in predicting the +ve class: HOTLEADS
- Target a value of 80%
- Check the model performance over the test data(confusion matrix,, Sensitivity, F1-score, ROC Curve
- Generate score variable

ROC CURVE


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.98)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
prob
0.0
0.2
0.4
0.6
0.8
1.0


0.0
0.2
0.4
0.6
0.8
1.0

CONCLUSION

- The most potential buyers are:
  - Tags_lost to EINS
  - Tags_Closed by Horizzon
  - Lead_Source_Welingak Website
- When the Lead Source was
  - Google
  - Direct traffic
  - Welingak Website
  - Organic Search
- When the last activity was
  - SMS
  - Olark chart conversion
- When the current occupation is “Working professional”
- Using these facts and figures X Education can reach their target of 80% and can convert more “Hot Leads”