Identifying Hotheads

Aryan

agenda

Business Problem

PRIMARY GOALS

Data Reading and Inspections

Data Preparation & EDA

Modeling and Optimization

Model summary

Model Evaluation

Finding Threshold

Summary

Business Problem

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

primary goals

There are quite a few goals for this case study:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

DATA READING AND INSPECTIONS

- There were **9240 leads** available in the data with **35 variables** to start analysing.
- 5 Numerical variables & 29 Categorical variables excluding Dependent variable
- Columns like **Prospect ID & Lead Number** are dropped as they were not useful
- Columns have missing value weather it is Category feature or Numerical Feature.
- For binary feature with Yes or No are replaced with 1 & 0
- For category variables, Missing value and Select Option are merged together to form single Missing class.
- Features namely **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, I agree to pay the amount through cheque** are dropped as they were 0 variance features. Discussed on next page.

Basic info for all variables available

Using data.info() Looking at the basic info of the columns such as:

- Count of variables
- Datatype of variable
- Non null values count to estimate missing i.e. Nan
- Incorrectly labelled variable

```
#   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Lead Origin                                   9240 non-null   object
 1   Lead Source                                   9240 non-null   object
 2   Do Not Email                                  9240 non-null   object
 3   Do Not Call                                   9240 non-null   object
 4   Converted                                     9240 non-null   int64
 5   TotalVisits                                   9103 non-null   float64
 6   Total Time Spent on Website                   9240 non-null   int64
 7   Page Views Per Visit                          9103 non-null   float64
 8   Last Activity                                 9240 non-null   object
 9   Country                                       9240 non-null   object
 10  Specialization                                9240 non-null   object
 11  How did you hear about X Education            9240 non-null   object
 12  What is your current occupation               9240 non-null   object
 13  What matters most to you in choosing a course 9240 non-null   object
 14  Search                                        9240 non-null   object
 15  Magazine                                      9240 non-null   object
 16  Newspaper Article                             9240 non-null   object
 17  X Education Forums                            9240 non-null   object
 18  Newspaper                                     9240 non-null   object
 19  Digital Advertisement                         9240 non-null   object
 20  Through Recommendations                       9240 non-null   object
 21  Receive More Updates About Our Courses        9240 non-null   object
 22  Tags                                          9240 non-null   object
 23  Lead Quality                                  9240 non-null   object
 24  Update me on Supply Chain Content             9240 non-null   object
 25  Get updates on DM Content                     9240 non-null   object
 26  Lead Profile                                  9240 non-null   object
 27  City                                          9240 non-null   object
 28  Asymmetrique Activity Index                   9240 non-null   object
 29  Asymmetrique Profile Index                    9240 non-null   object
 30  Asymmetrique Activity Score                   5022 non-null   float64
 31  Asymmetrique Profile Score                    5022 non-null   float64
 32  I agree to pay the amount through cheque      9240 non-null   object
 33  A free copy of Mastering The Interview        9240 non-null   object
 34  Last Notable Activity                         9240 non-null   object
```

DATA PREPARATION & EDA

- Encoding
- Dropping redundant
- Merging less used Classes
- Dealing with missing data

Calculating Variable's nulls count

Code used:

- `[print(i,data[i].isna().sum())for i in data.columns if data[i].isna().sum()>0]`

- Above code is used to find the features that have the missing values in them.
- Most misisng data was in Lead Quality

- Also *Asymmetrique Activity Index,Asymmetrique Profile Index,Asymmetrique Activity Score, Asymmetrique Profile Score* all have same missing value.
- Variables with missing value percentage more than 40% are dropped.

Variables	Count
Lead Source	0036
TotalVisits	0137
Page Views Per Visit	0137
Last Activity	0103
Country	2461
Specialization	1438
How did you hear about X Education	2207
What is your current occupation	2690
What matters most to you in choosing a course	2709
Tags	3353
Lead Quality	4767
Lead Profile	2709
City	1420
Asymmetrique Activity Index	4218
Asymmetrique Profile Index	4218
Asymmetrique Activity Score	4218
Asymmetrique Profile Score	4218

Dropping 0 variance Features

```
Do Not Call Do Not Call
0    0.999784
1    0.000216
Name: proportion, dtype: float64
Search Search
0    0.998485
1    0.001515
Name: proportion, dtype: float64
Magazine Magazine
0    1.0
Name: proportion, dtype: float64
Newspaper Article Newspaper Article
0    0.999784
1    0.000216
Name: proportion, dtype: float64
X Education Forums X Education Forums
0    0.999892
1    0.000108
Name: proportion, dtype: float64
Newspaper Newspaper
0    0.999892
1    0.000108
Name: proportion, dtype: float64
Digital Advertisement Digital Advertisement
0    0.999567
1    0.000433
```

These variables are dropped due to zero variance i.e. one class is present 99%.
Refer to attached images for more information

```
Digital Advertisement Digital Advertisement
0    0.999567
1    0.000433
Name: proportion, dtype: float64
Through Recommendations Through Recommendations
0    0.999242
1    0.000758
Name: proportion, dtype: float64
Receive More Updates About Our Courses Receive More Updates About Our Courses
0    1.0
Name: proportion, dtype: float64
Update me on Supply Chain Content Update me on Supply Chain Content
0    1.0
Name: proportion, dtype: float64
Get updates on DM Content Get updates on DM Content
0    1.0
Name: proportion, dtype: float64
I agree to pay the amount through cheque I agree to pay the amount through cheque
0    1.0
```

Other Categorical Variables dropped

Variables that were dropped other than 0 variance features. Variables that are dropped along with the reason are mentioned in the table

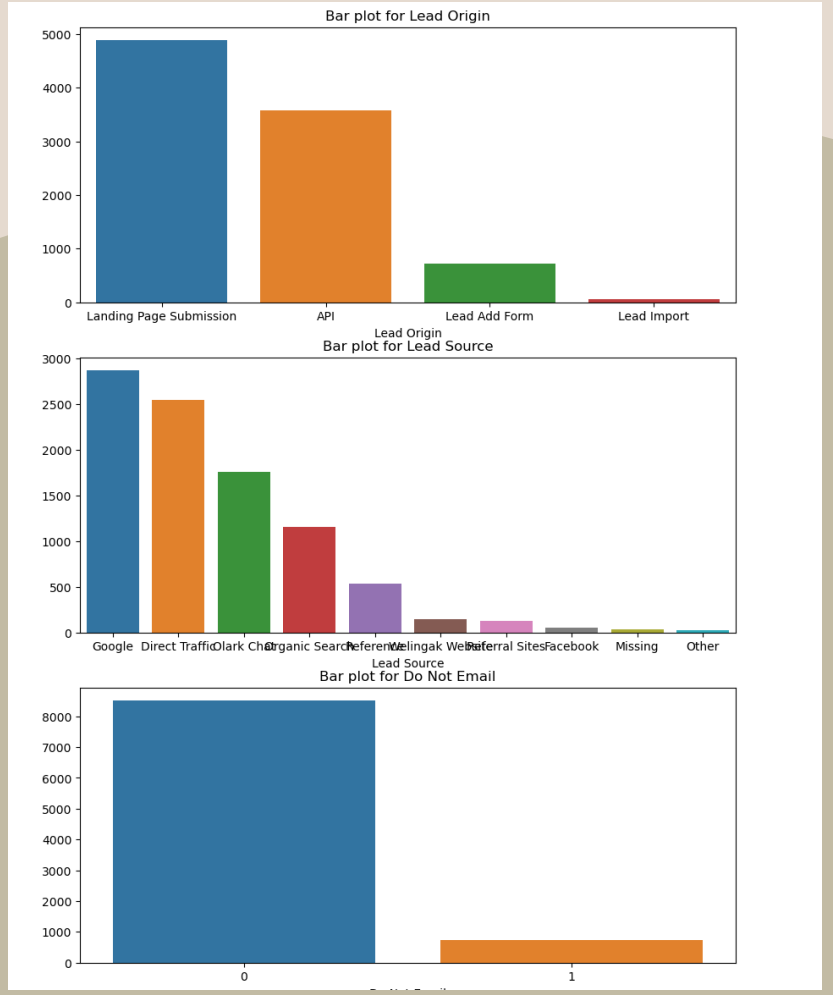
Variables	Reason
Prospect ID	Redundant data
Lead Number	Redundant data
Lead Quality	Dropped because feature is based on employee intuitions
Asymmetrique Activity Index	Dropped due to high percentage of missing values
Asymmetrique Profile Index	Dropped due to high percentage of missing values
Lead Profile	High percent data missing
How did you hear about X Education	Dropped due to high missing value percentage
Country	Due to 95 percent data is belong to India or missing

Univariate Analysis Action

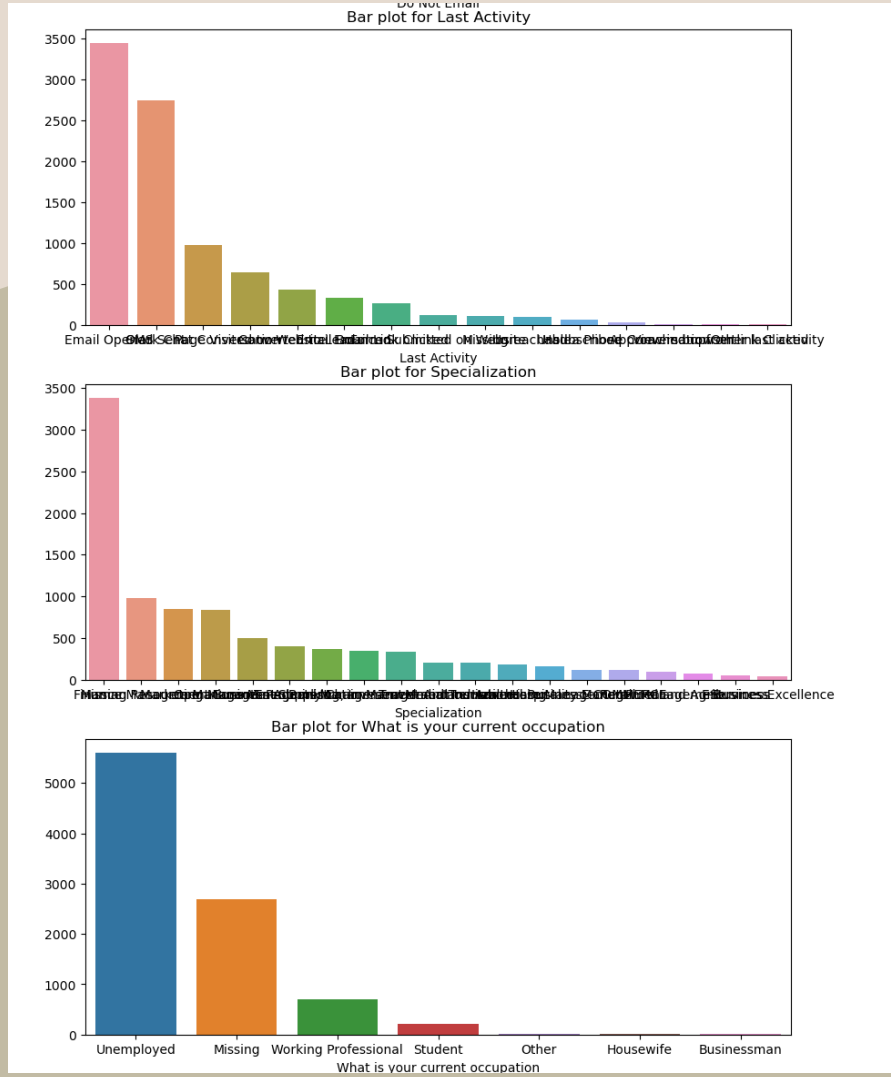
- In Lead Origin variable **`Quick Add Form`** has single count so it is merged with **`Lead Import`**

In Lead source all the classes having count smaller than **5** i.e. **bing,,google,,Click2call, Press_Release, Social Media, Live Chat, youtubechannel, testone, Pay per Click Ads, welearnblog_Home, WeLearn, blog, NC_EDM** are merged to make single class **Other**

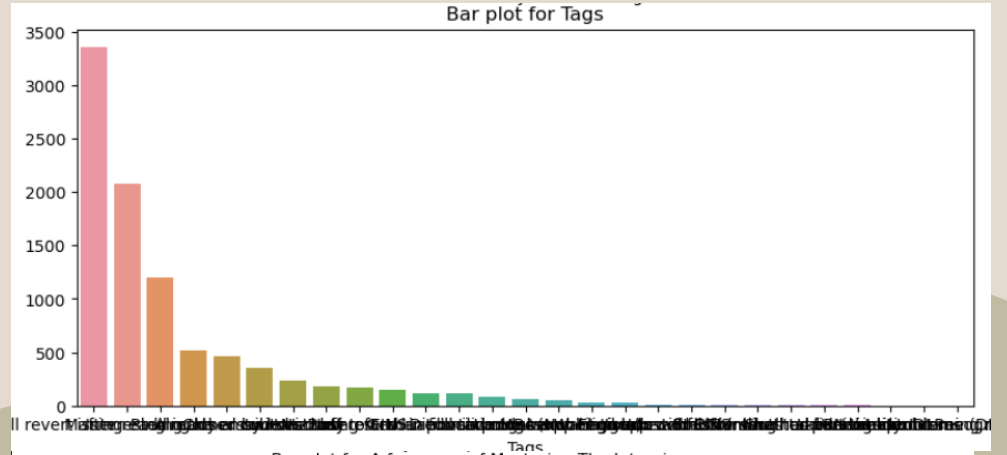
- In **Do not email** Yes and no are converted into binary number 1 & 0
- Graphs shown are post action applied

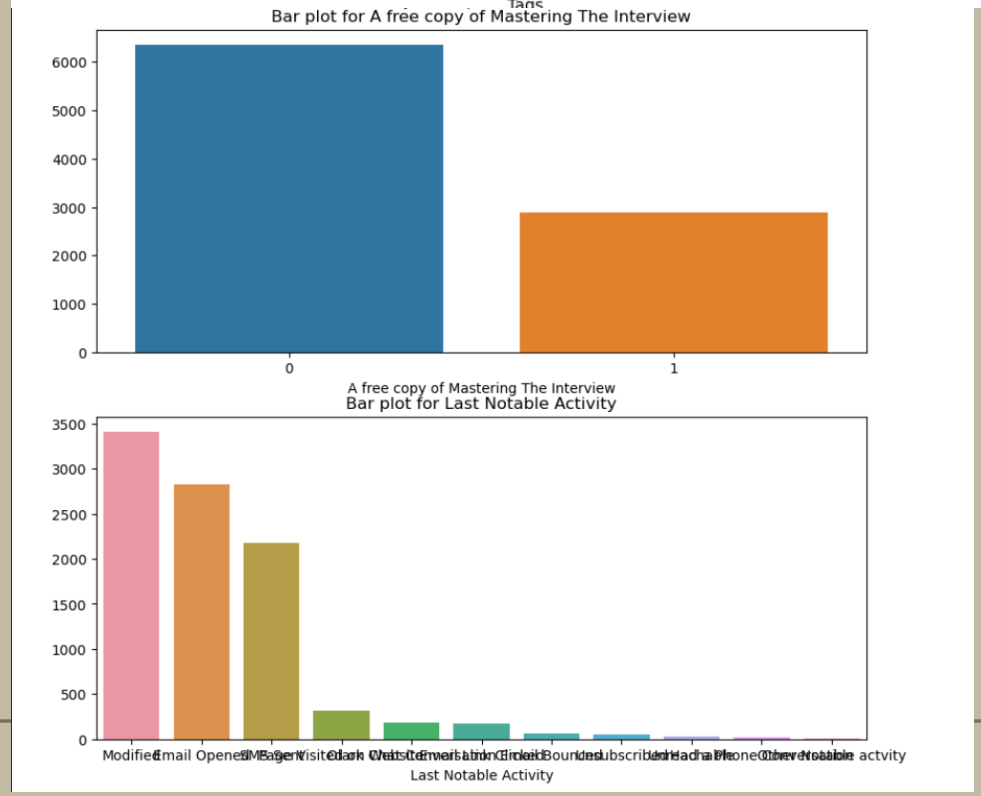


- In last Activity variables **Email Opened, SMS Sent, Olark Chat Conversation, Page Visited on Website, Converted to Lead, Email Bounced, Email Link Clicked, Form Submitted on Website, Missing, Unreachable, Unsubscribed, Had a Phone Conversation, Approached upfront, View in browser link Clicked, Email Received, Email Marked Spam, Visited Booth in Tradeshow, Resubscribed to emails** are merged into 1 class **Other last activity**
- For **Specilization** Variable all Select are merged as **Missing** Class

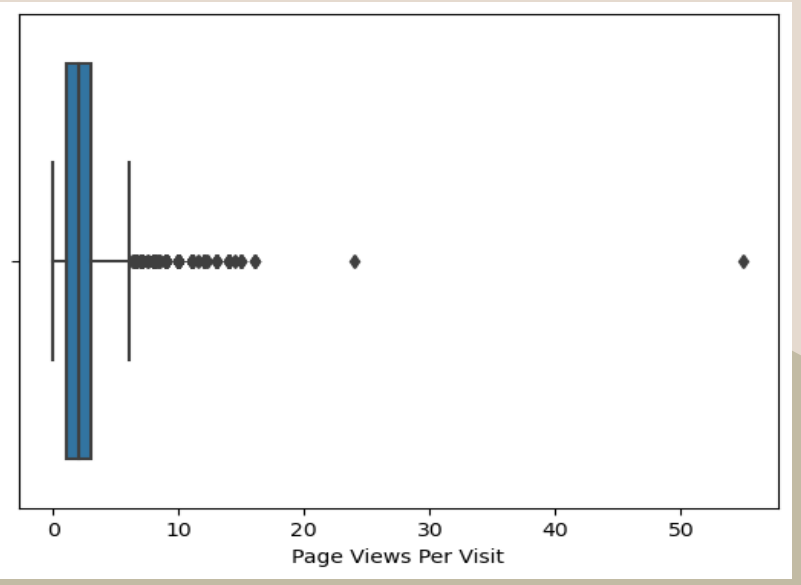


- Variable **A free copy of Mastering the** Interview was converted into 0 & 1 for respective NO and Yes.
- In **Last Notable Activity** the classes **Email Marked Spam**, **Approached upfront**, **Resubscribed to emails**, **View in browser link Clicked**, **Form Submitted on Website**, **Email Receive** are merged into **`Other Notable activity`** due to less count

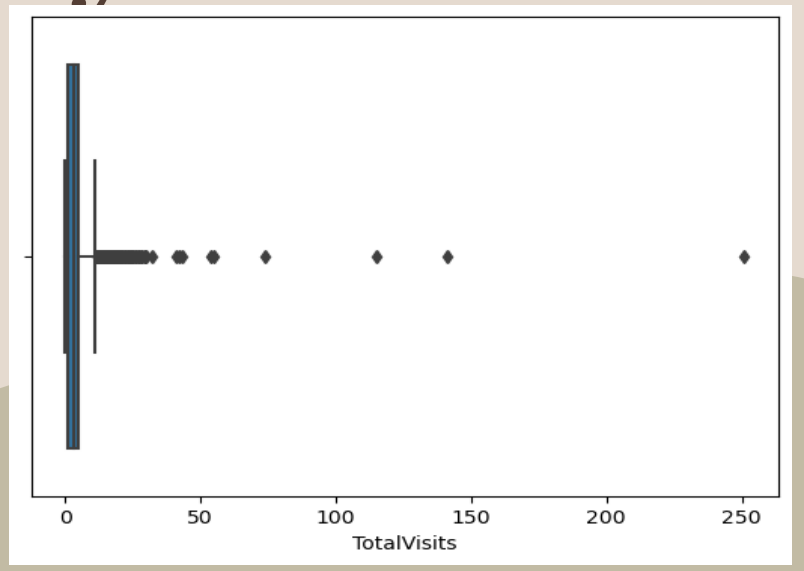




Numerical Variables analysis



Page Views Per Visit		TotalVisits
9103.000000	count	9103.000000
2.362820	mean	3.445238
2.161418	std	4.854853
0.000000	min	0.000000
2.000000	50%	3.000000
4.000000	80%	5.000000
5.000000	90%	7.000000
6.000000	95%	10.000000
9.000000	99%	17.000000
55.000000	max	251.000000



- Per views per visit also appears to have outliers 55.
- When observed carefully there is only 5 cases having Page per visit more than 15.
- This can be due to curiosity of Person
- Missing values are filled with Median values

- Variable TotalVisits appear to have outliers 251
- But on further investigation 99% have 17 TotalVisit while only 10 cases of visit more than 30.
- This can be the case that people either waiting for price drop or any offer or for specific course
- Missing values are filled with Median values

Encoding Category and feature scaling

- Category variables `Lead Origin, Lead Source, Do Not Email, Last Activity, Specialization, What is your current occupation, What matters most to you in choosing a course, Tags, City, A free copy of Mastering The Interview , Last Notable Activity` are encoded using `pandas.get_dummies()`
- `And Numerical Columns Total Time Spent on Website, TotalVisits, Page Views Per Visit` are Standardised using StandardScaler()

Model creation and Optimization

Model Summary

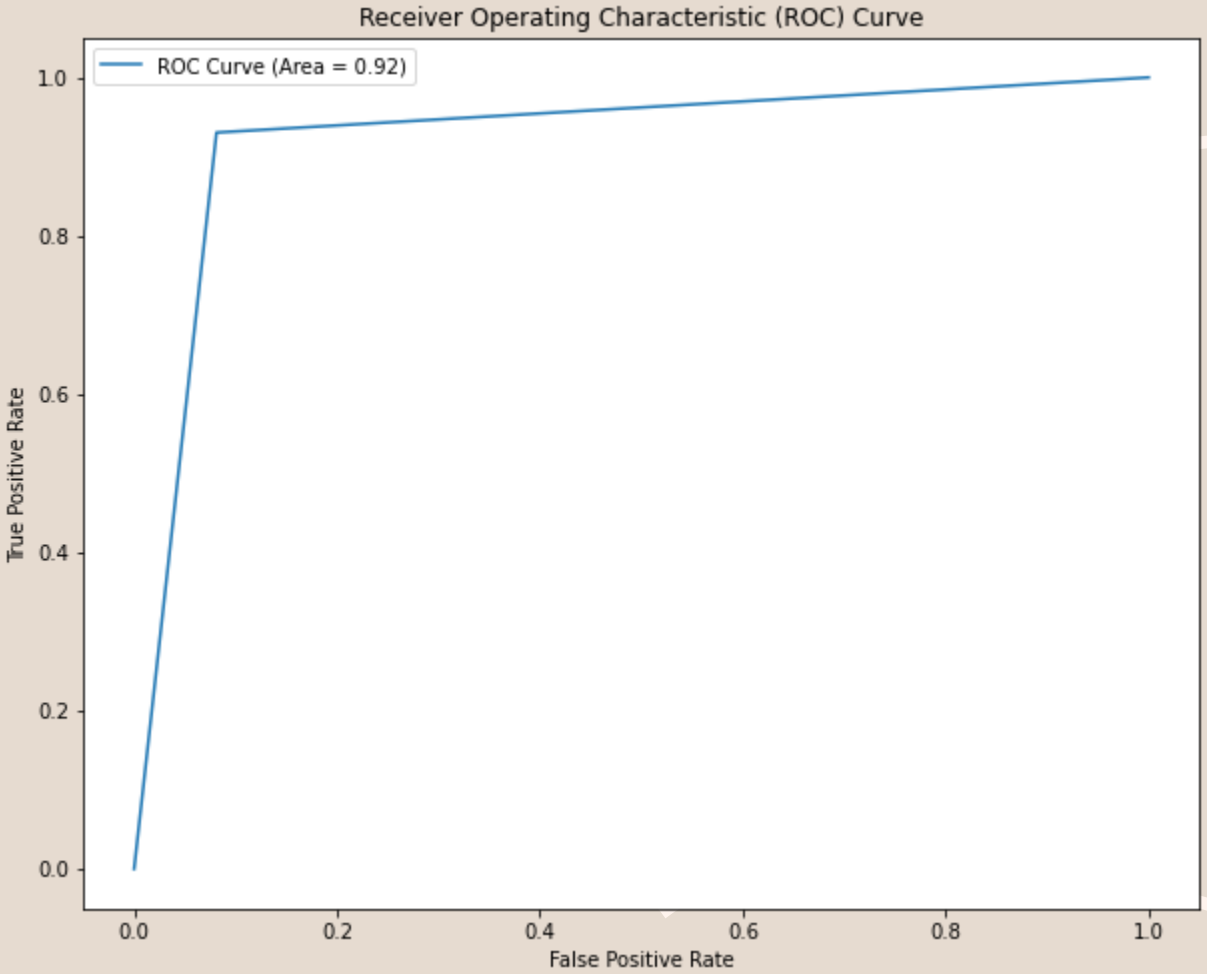
Multiple models are made and tested and optimised for better efficiency and reduced Variables using technique such as RFE, VIF and Custom fine tuning using Model Summary.

Important Features for Model

- Tags
- Lead Source
- Last Notable activity

Features	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.4192	0.156	-28.365	0.000	-4.725	-4.114
Lead Source_Welingak Website	3.5935	1.018	3.530	0.000	1.598	5.589
Tags_Busy	3.6010	0.252	14.262	0.000	3.106	4.096
Tags_Closed by Horizzon	9.9022	1.014	9.764	0.000	7.915	11.890
Tags_Lost to EINS	10.6054	1.030	10.296	0.000	8.587	12.624
Tags_Missing	4.4575	0.186	23.947	0.000	4.093	4.822
Tags_Will revert after reading the email	7.2919	0.216	33.698	0.000	6.868	7.716
Tags_switched off	-1.7262	0.728	-2.370	0.018	-3.154	-0.298
Lead Profile_Missing	-2.4753	0.138	-17.993	0.000	-2.745	-2.206
Last Notable Activity_SMS Sent	2.5453	0.124	20.478	0.000	2.302	2.789

ROC curve and model Evaluations



Model is good it have AUC = 0.91 indicating good model

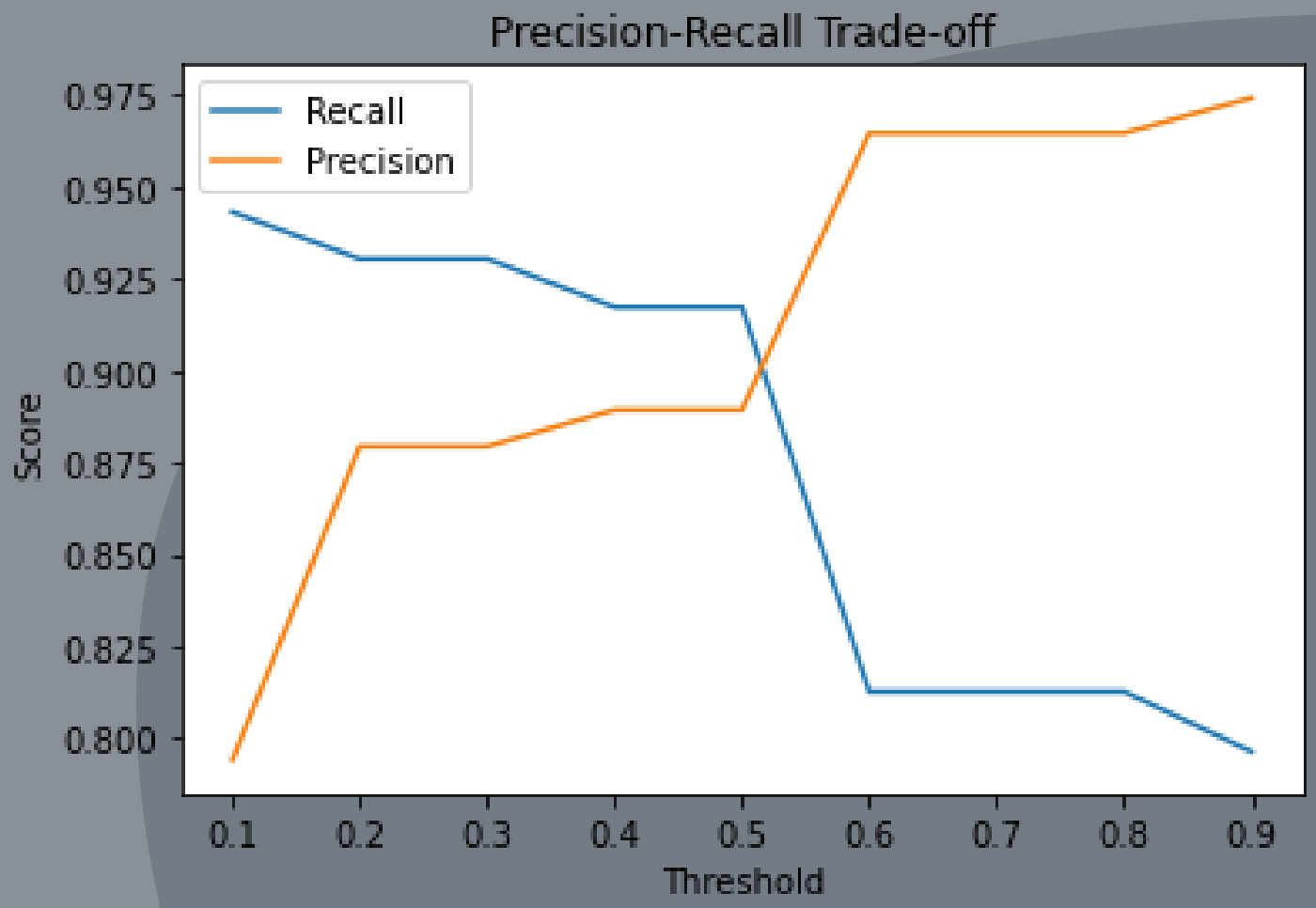
Model evaluation on training data:

Recall Score:	92%
Precision Score	89%
Accuracy	92%
Specificity	93%
False Negative Rate	08%

Model Evaluation for test data:

Recall Score	93%
Precision Score	88%
Accuracy	92%
Specificity	92%
False Negative Rate	07%

Finding Threshold



Finding the optimal value i.e. Threshold

- Threshold value is obtained by plotting various value of Recall score and Precision score on the same plot against values of threshold.
- Threshold value find out to be 0.505556
- Or on other word any lead with score more than 51 is a potential **Hot Lead**

summary

- The model provide Score to each lead on a scale of 0 to 100 higher the number more is the probability of getting into converted.
- Any lead with a Lead score more than 51(threshold) can be considered as Hot Lead.
- Model have a accuracy of 92% and a average false positive rate of 7.5 %.

thank you

Aryan Dhiman

dhimanarya650@gmail.com