LEAD SCORING ASSIGNMENT

1. Importing and Understanding Data

We import the necessary library and look at the information of data set

- Size: 9240 entries x 37 columns
- There are 4 numeric variables
- The variables with high rate of null values: *'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'*, *'Prospect ID'*, *'Lead Number'*, *'Tags'*, *'Lead Quality'*, *'Lead Profile'*, *'Asymmetrique Activity Index'*, *'Asymmetrique Profile Index'*, *'Asymmetrique Activity Score'*, *'Asymmetrique Profile Score'*, *'Magazine'*.

```
 #   Column                                         Non-Null Count  Dtype
---  ------                                         --------------  -----
 0   Prospect ID                                    9240 non-null   object
 1   Lead Number                                    9240 non-null   int64
 2   Lead Origin                                    9240 non-null   object
 3   Lead Source                                    9204 non-null   object
 4   Do Not Email                                   9240 non-null   object
 5   Do Not Call                                    9240 non-null   object
 6   Converted                                      9240 non-null   int64
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object
 11  Country                                        6779 non-null   object
 12  Specialization                                 7802 non-null   object
 13  How did you hear about X Education             7033 non-null   object
 14  What is your current occupation                6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                         9240 non-null   object
 17  Magazine                                       9240 non-null   object
 18  Newspaper Article                              9240 non-null   object
 19  X Education Forums                             9240 non-null   object
 20  Newspaper                                      9240 non-null   object
 21  Digital Advertisement                          9240 non-null   object
 22  Through Recommendations                        9240 non-null   object
 23  Receive More Updates About Our Courses         9240 non-null   object
 24  Tags                                           5887 non-null   object
 25  Lead Quality                                   4473 non-null   object
 26  Update me on Supply Chain Content              9240 non-null   object
 27  Get updates on DM Content                      9240 non-null   object
 28  Lead Profile                                   6531 non-null   object
 29  City                                           7820 non-null   object
 30  Asymmetrique Activity Index                    5022 non-null   object
 31  Asymmetrique Profile Index                     5022 non-null   object
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object
 35  A free copy of Mastering The Interview         9240 non-null   object
 36  Last Notable Activity                          9240 non-null   object
dtypes: float64(4), int64(3), object(30)
```

2. Data Preparation

Combine variables into 3 groups for processing:

drop_var_list = ['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','Prospect ID','Lead Number','Tags','Lead Quality','Lead Profile','Asymmetrique Activity Index','Asymmetrique Profile Index','Asymmetrique Activity Score','Asymmetrique Profile Score','Magazine']

**We drop the variable that has a high rate of null values.*

get_dummies_vars = ['What is your current occupation','What matters most to you in choosing a course','Country','Lead Origin','Lead Source','Last Activity','Specialization','How did you hear about X Education','City','Last Notable Activity']

binary_vars = ['Do Not Email','Do Not Call','Search','Newspaper Article','X Education Forums','Newspaper','Digital Advertisement','Through Recommendations','A free copy of Mastering The Interview']

num_vars = ['TotalVisits','Total Time Spent on Website','Page Views Per Visit']

2. Data Preparation

We handle the categorical variables *'What is your current occupation','What matters most to you in choosing a course','Country','Lead Origin','Lead Source','Last Activity','Specialization','How did you hear about X Education','City','Last Notable Activity‘* using dummy variables.

Finally, we add the dummy variable tables *lead_origin_dummies,lead_source_dummies,last_activity_dummies,last_notable_activity_dummies,specialization_dummies,city_dummies to the Leads dataset after drop the imbalanced variables*

2. Data Preparation

Splitting the Data into Training and Testing Sets

```
df_train, df_test = train_test_split(Leads, train_size = 0.7, test_size = 0.3, random_state = 100)
```

Rescaling the Features using MinMaxScaler

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df_train[num_vars] = scaler.fit_transform(df_train[num_vars])
```

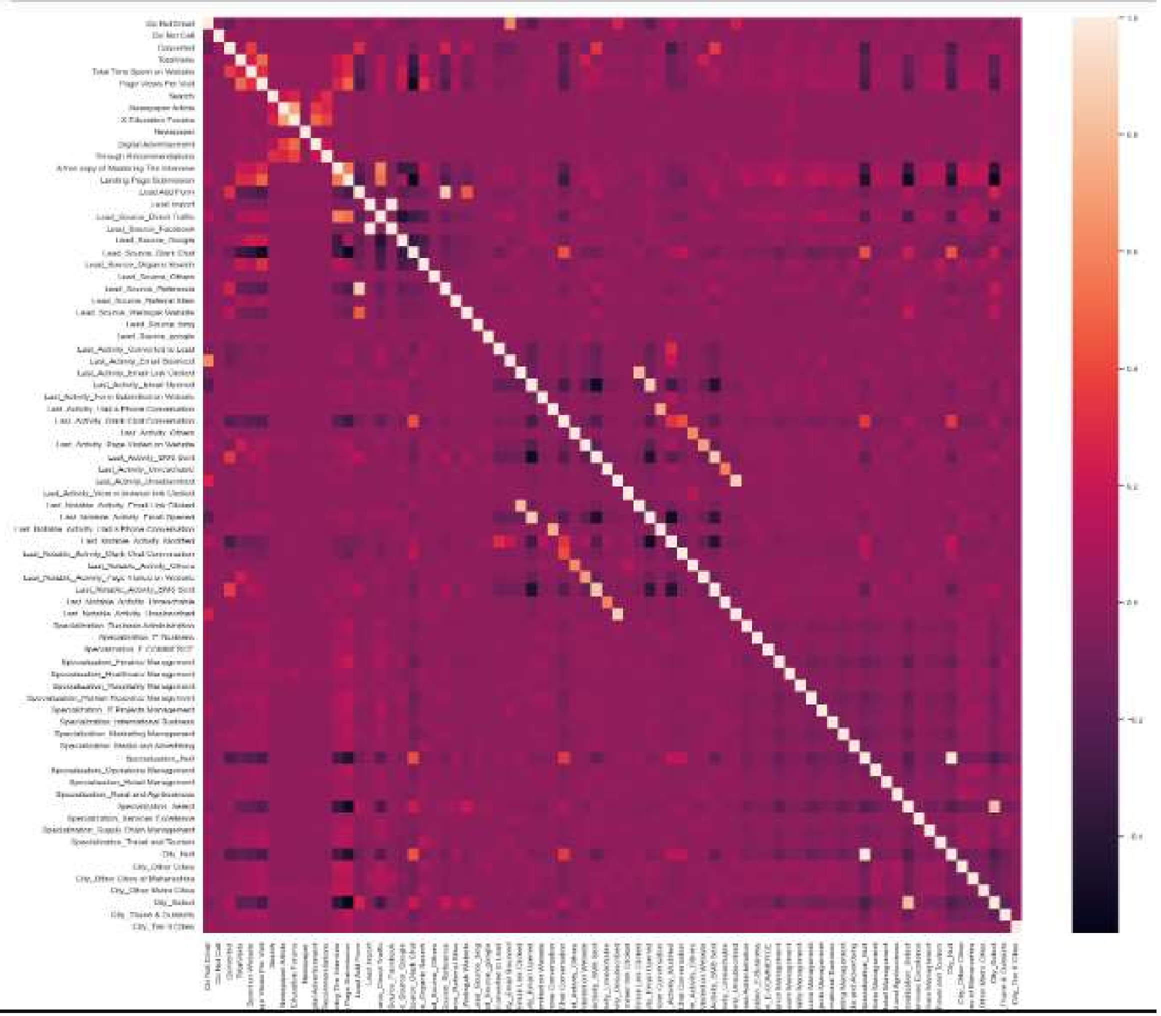
Dividing into X and Y sets for the model building

```
y_train = df_train.pop('converted')
X_train = df_train
```

2. Data Preparation

Looking at Correlations

- We remove some highly correlated variables
- Those are *'Lead Add Form','Lead Import','Specialization_Null','Specialization_Select','Last_Activity_Email Opened','Last_Activity_SMS Sent','Last_Activity_Unsubscribed*



3. Building the Logistic model

RFE

Importing RFE and LogisticRegression

```
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
```

Running RFE with the output number of the variable equal to 20

```
logreg = LogisticRegression()
rfe = RFE(logreg,n_features_to_select=20)
rfe = rfe.fit(X_train, y_train)
```

3. Building the Logistic model

RFE

Columns that we keep after using RFE:

```
Index(['Do Not Email', 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit',
       'Lead_Source_Direct Traffic', 'Lead_Source_Google', 'Lead_Source_Organic Search',
       'Lead_Source_Reference', 'Lead_Source_Referral Sites', 'Lead_Source_Welingak Website',
       'Lead_Source_google', 'Last_Activity_Email Bounced', 'Last_Activity_Had a Phone Conversation',
       'Last_Activity_Olark Chat Conversation', 'Last_Notable_Activity_Email Link Clicked',
       'Last_Notable_Activity_Email Opened', 'Last_Notable_Activity_Modified',
       'Last_Notable_Activity_Olark Chat Conversation', 'Last_Notable_Activity_Page Visited on Website',
       'City_Null'], dtype='object')
```

3. Building the Logistic model

Building model using statsmodel, for the detailed statistics

- Creating X_test dataframe with RFE selected variables
- Adding a constant variable
- Running the linear model using GLM method

3. Building the Logistic model

We drop the variables that have high P-value

The summary of our Logistic model =>

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6332
Model Family:	Binomial	Df Model:	18
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2726.3
Date:	Fri, 25 Aug 2023	Deviance:	5452.6
Time:	00:12:49	Pearson chi2:	6.56e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3778
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9374	0.106	8.876	0.000	0.730	1.144
Do Not Email	-1.5709	0.194	-8.079	0.000	-1.952	-1.190
TotalVisits	8.9850	2.479	3.624	0.000	4.125	13.845
Total Time Spent on Website	4.6250	0.161	28.638	0.000	4.309	4.942
Page Views Per Visit	-1.6100	0.555	-2.902	0.004	-2.697	-0.522
Lead_Source_Direct Traffic	-1.7354	0.130	-13.343	0.000	-1.990	-1.480
Lead_Source_Google	-1.2557	0.126	-9.963	0.000	-1.503	-1.009
Lead_Source_Organic Search	-1.4324	0.156	-9.195	0.000	-1.738	-1.127
Lead_Source_Reference	2.7040	0.232	11.655	0.000	2.249	3.159
Lead_Source_Referral Sites	-1.3825	0.337	-4.108	0.000	-2.042	-0.723
Lead_Source_Welingak Website	4.2724	0.731	5.846	0.000	2.840	5.705
Last_Activity_Email Bounced	-1.2687	0.420	-3.021	0.003	-2.092	-0.446
Last_Activity_Olark Chat Conversation	-1.0904	0.189	-5.775	0.000	-1.460	-0.720
Last_Notable_Activity_Email Link Clicked	-1.7867	0.253	-7.055	0.000	-2.283	-1.290
Last_Notable_Activity_Email Opened	-1.4017	0.086	-16.233	0.000	-1.571	-1.232
Last_Notable_Activity_Modified	-1.8725	0.095	-19.675	0.000	-2.059	-1.686
Last_Notable_Activity_Olark Chat Conversation	-1.6597	0.376	-4.417	0.000	-2.396	-0.923
Last_Notable_Activity_Page Visited on Website	-1.8764	0.207	-9.081	0.000	-2.281	-1.471
City_Null	-1.3210	0.127	-10.380	0.000	-1.570	-1.072

3. Building the Logistic model

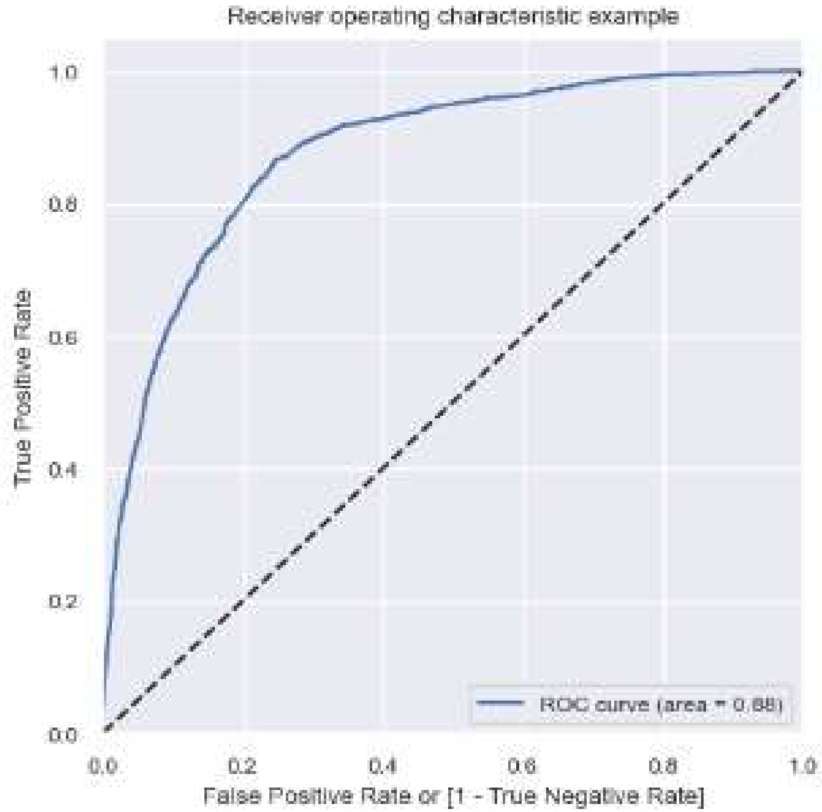
Checking and dealing with multicollinearity

- All variables have a good value of VIF. So no need to drop any variables and we can proceed with making predictions using this model only

	Features	VIF
3	Page Views Per Visit	4.56
5	Lead_Source_Google	3.39
4	Lead_Source_Direct Traffic	3.00
14	Last_Notable_Activity_Modified	2.43
6	Lead_Source_Organic Search	2.32
2	Total Time Spent on Website	2.29
1	TotalVisits	2.01
0	Do Not Email	1.85
11	Last_Activity_Olark Chat Conversation	1.84
13	Last_Notable_Activity_Email Opened	1.77
10	Last_Activity_Email Bounced	1.77
17	City_Null	1.41
15	Last_Notable_Activity_Olark Chat Conversation	1.35
16	Last_Notable_Activity_Page Visited on Website	1.18
8	Lead_Source_Referral Sites	1.15
7	Lead_Source_Reference	1.07
12	Last_Notable_Activity_Email Link Clicked	1.05
9	Lead_Source_Welingak Website	1.02

4. Model Evaluation

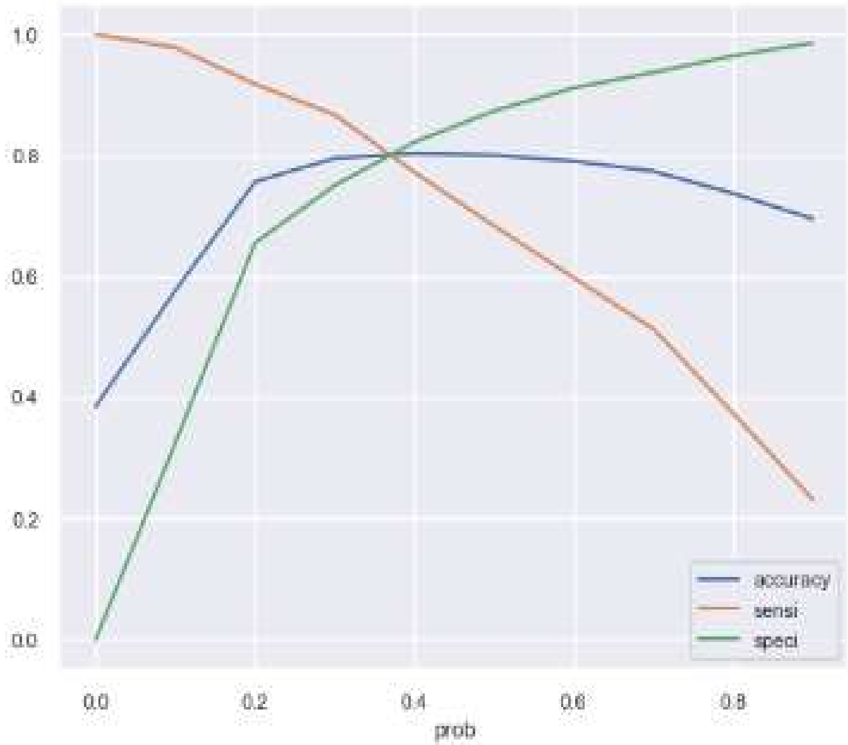
Plotting the ROC Curve



4. Model Evaluation

Finding Optimal Cutoff Point

From the curve, 0.37 is the optimum point to take it as a cutoff probability.



4. Model Evaluation

Metric scoring by Optimal Cutoff Point

```
metrics.accuracy_score(y_train_pred_final.Churn, y_train_pred_final.predicted)
```

0.8009762242166588

```
metrics.precision_score(y_train_pred_final.Churn, y_train_pred_final.predicted)
```

0.7075140449438202

```
metrics.recall_score(y_train_pred_final.Churn, y_train_pred_final.predicted)
```

0.8237939493049877

```
metrics.f1_score(y_train_pred_final.Churn, y_train_pred_final.predicted)
```

0.7612391386475255

Precision

TP / TP + FP

```
confusion[1,1]/(confusion[0,1]+confusion[1,1])
```

0.7075140449438202

Recall

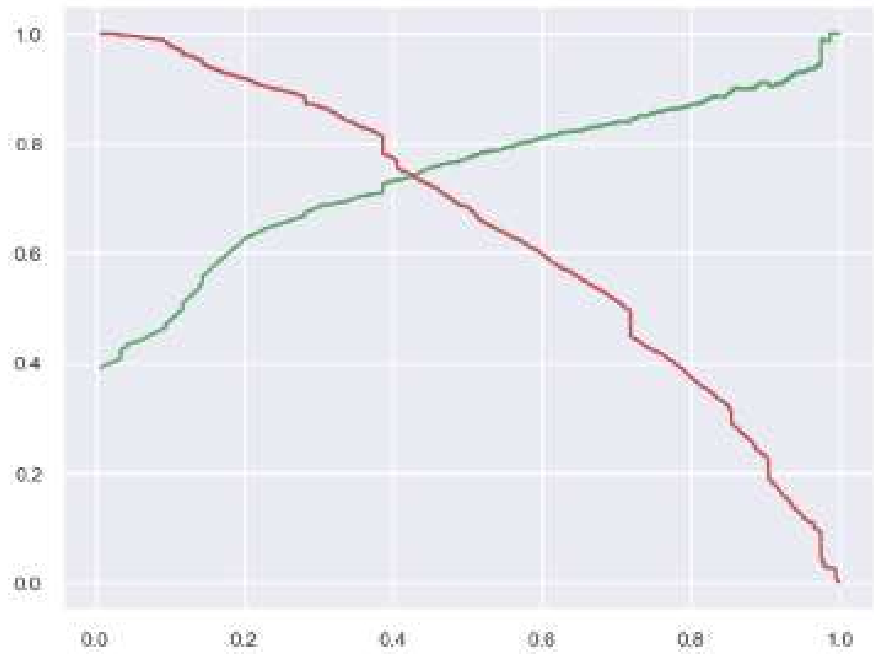
TP / TP + FN

```
confusion[1,1]/(confusion[1,0]+confusion[1,1])
```

0.8237939493049877

4. Model Evaluation

Precision and recall tradeoff



4. Model Evaluation

Making predictions on the test set

```
y_pred_final['final_predicted'] = y_pred_final.Churn_Prob.map(lambda x: 1 if x > 0.43 else 0)
```

```
# Let's check the overall accuracy.
metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
0.7848616966580977
```

```
# Let's see the sensitivity of our logistic regression model
TP / float(TP+FN)
```

```
0.7836198179979778
```

```
# Let us calculate specificity
TN / float(TN+FP)
```

```
0.7843137254901961
```

Summary

After data preparation, model building and evaluation, we get a logistic regression model of variables that affect the the ability to convert for potential customers including:

- Variables that have the positive effect:
  + *TotalVisits*: Customers with more total visits are more likely to convert.
  + *Total Time Spent on Website*: waCustomers with more total time on the website are more likely to convert.
  + Lead_Source's dummy: *Lead_Source_Reference, Lead_Source_Welingak Website* : Customers with Lead Source as Reference and Welingak Website are more likely to convert.
- Variables that have the opposite effect:
  + *Do Not Email*: Customers requesting no email are less likely to convert
  + *Page Views Per Visit*: Customers with high Page Views Per Visit are less likely to convert.
  + Lead_Source's dummy: *Lead_Source_Direct Traffic, Lead_Source_Google, Lead_Source_Organic Search, Lead_Source_Referral Sites*: Customers with Lead Source as Direct Traffic, Google, Organic Search and Referral Sites are less likely to convert.
  + Last_Activity's dummy: *Last_Activity_Email Bounced, Last_Activity_Olark Chat Conversation:* Customers with last activity as Email Bounced, Olark Chat Conversation are less likely to convert.
  + Last_Notable_Activity's dummy: *Last_Notable_Activity_Email Link Clicked, Last_Notable_Activity_Email Opened, Last_Notable_Activity_Modified, Last_Notable_Activity_Olark Chat Conversation, Last_Notable_Activity_Page Visited on Website* : Customers with last notable activity as Email Link Clicked, Email Opened, Modified, Olark Chat Conversation, Page Visited on Website are less likely to convert.
  + *City_Null*: Customers without city information are less likely to convert

Summary

- Top three variables in your model which contribute most towards the probability of a lead getting converted with their coefficients:
    - *TotalVisits* 8.985
    - *Total Time Spent on Website* 4.625
    - *Lead_Source_Welingak Website* 4.272

- Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion with their coefficients:
    - *Lead_Source_Welingak Website* 4.272
    - *Lead_Source_Reference* 2.704
    - *Last_Notable_Activity_Page Visited on Website* -1.876