Lead scoring Assignment

Manjot singh

Problem Statement

X Education, an online education company, receives daily website traffic from industry professionals interested in their courses. The company markets its courses on various platforms, attracting leads through form submissions, video views, and referrals. However, their lead conversion rate is low, with only 30% successfully converting. To improve efficiency, X Education aims to identify the most promising leads, or 'Hot Leads,' allowing the sales team to focus on them for increased conversion rates. Their goal is an 80% lead conversion rate. As an appointed consultant, your task is to develop a lead scoring model, prioritizing leads with higher scores for improved conversion chances.

Data Analysis/Preparation

Conversion of Binary YES/NO to Numeric 0-1

```
# List of variables to map

varlist =  ['Do Not Email' ,'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',

# Defining the map function
def binary_map(x):
    return x.map({'Yes': 1, "No": 0})

# Applying the function to the housing list
leads_data[varlist] = leads_data[varlist].apply(binary_map)
```

Handling of special value ‘Select’ in certain columns

```
leads_data['Specialization'] = leads_data['Specialization'].apply(lambda x : (None if x == 'Select' else x))
leads_data['How did you hear about X Education'] = leads_data['How did you hear about X Education'].apply(lambda x : (N
leads_data['Lead Profile'] = leads_data['Lead Profile'].apply(lambda x : (None if x == 'Select' else x))
leads_data['City'] = leads_data['City'].apply(lambda x : (None if x == 'Select' else x))
```

Data Analysis/Preparation

Removed the column with null% more than 50%

```
leads_data.isnull().sum()
```

```
Prospect ID                                       0
Lead Number                                       0
Lead Origin                                       0
Lead Source                                      36
Do Not Email                                      0
Do Not Call                                       0
Converted                                         0
TotalVisits                                     137
Total Time Spent on Website                       0
Page Views Per Visit                            137
Last Activity                                   103
Country                                        2461
Specialization                                 3380
How did you hear about X Education             7250
What is your current occupation                2690
What matters most to you in choosing a course  2709
Search                                            0
Magazine                                          0
Newspaper Article                                 0
X Education Forums                                0
Newspaper                                         0
Digital Advertisement                             0
Through Recommendations                           0
Receive More Updates About Our Courses            0
Tags                                           3353
Lead Quality                                   4767
Update me on Supply Chain Content                 0
Get updates on DM Content                         0
Lead Profile                                   6855
City                                           3669
Asymmetrique Activity Index                    4218
Asymmetrique Profile Index                     4218
Asymmetrique Activity Score                    4218
Asymmetrique Profile Score                     4218
I agree to pay the amount through cheque          0
A free copy of Mastering The Interview            0
Last Notable Activity                             0
dtype: int64
```

Did the grouping of categorical variables before creating the dummy variables

```
def handle_leads(x):
    if x in ('Quick Add Form','Lead Add Form','Lead Import'):
        return 'LessFrequent'
    else:
        return x

leads_data['Lead Origin'] = leads_data['Lead Origin'].apply(handle_leads)

def handle_lead_source(x):
        if x in ['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']:
            return 'LessFrequent'
        #'Referral Sites', 'Reference'
        else:
            return x

leads_data['Lead Source'] = leads_data['Lead Source'].apply(handle_lead_source)

def handle_Last_Notable_Activity(x):
    #Olark Chat Conversation
        if x in ['Approached upfront','Email Bounced','Email Link Clicked','Email Marked Spam','Email Received','Form S
            return 'LessFrequent'
        else:
            return x

leads_data['Last Notable Activity'] = leads_data['Last Notable Activity'].apply(handle_Last_Notable_Activity)

def handle_last_Activity(x):
        if x in ['Email Marked Spam','Email Received','Form Submitted on Website','Had a Phone Conversation','Resubscri
                 'Visited Booth in Tradeshow']:
            return 'LessFrequent'
        else:
            return x

leads_data['Last Activity'] = leads_data['Last Activity'].apply(handle_last_Activity)
```

Feature Scaling

Did the Feature scaling for variables having high numeric values

Feature Scaling

```
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()

X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']] = scaler.fit_transform(X_train[['TotalVis

X_train.head()
X_train = X_train.fillna(X_train.mean())
X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']].describe()
```

Model Building

After Multiple iterations and removing columns bases on VIF and P values. We got the model accuracy of **76**% and these are the final results

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6391	0.109	-5.886	0.000	-0.852	-0.426
Do Not Email	-1.4224	0.168	-8.466	0.000	-1.752	-1.093
Total Time Spent on Website	0.7702	0.034	22.974	0.000	0.705	0.836
Lead Source_Google	-0.2454	0.069	-3.557	0.000	-0.381	-0.110
Lead Source_Olark Chat	0.3105	0.095	3.274	0.001	0.125	0.496
Last Activity_Converted to Lead	-1.3382	0.208	-6.428	0.000	-1.746	-0.930
Last Activity_Email Bounced	-1.4190	0.388	-3.660	0.000	-2.179	-0.659
Last Activity_Email Link Clicked	-1.0321	0.252	-4.095	0.000	-1.526	-0.538
Last Activity_Email Opened	0.1169	0.105	1.117	0.264	-0.088	0.322
Last Activity_Olark Chat Conversation	-1.3765	0.163	-8.428	0.000	-1.697	-1.056
Last Notable Activity_LessFrequent	1.1298	0.236	4.779	0.000	0.666	1.593
Last Notable Activity_Modified	-0.1462	0.095	-1.545	0.122	-0.332	0.039
Last Notable Activity_SMS Sent	1.5660	0.120	13.010	0.000	1.330	1.802

COEF and P-values

```
In [173]: y_pred_final['final_predicted'] = y_pred_final.Conversion_Prob.map(lambda x: 1
```

```
In [174]: y_pred_final.head()
```

Out[174]:

	Converted	CustID	Conversion_Prob	final_predicted
0	0	3271	0.226150	0
1	1	1490	0.717374	1
2	0	7936	0.206284	0
3	1	4216	0.205405	0
4	0	3830	0.273905	0

```
In [175]: metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted)
```

```
Out[175]: 0.7631289019463827
```

```
In [ ]:
```

Model accuracy