LEAD SCORING CASE STUDY

Logistic Regression Model

PROBLEM STATEMENT

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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. 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 lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

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%.

APPROACH

Below need to be done to ensure we reach to a model which can help us to meet the business requirement

Steps to be performed

1. *Importing all the required libraries*
2. *Reading and understanding the Dataset*
3. *Data preprocessing & EDA*
4. *Model building & Feature selection (Using RFE)*
5. *Model Evaluation*
6. *Conclusion/Summary*

IMPORTING ALL THE REQUIRED LIBRARIES

jupyter LeadScoringCaseStudy Last Checkpoint: 9 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

1.Importing all the required libraries

In [1]:
```
# importing Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import RFE
from sklearn.metrics import r2_score , mean_squared_error
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.preprocessing import MinMaxScaler
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.model_selection import cross_val_score
```

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
from sklearn.model_selection import cross_val_score
```

In [2]:
```
# to suppress warning importing libraries
import warnings
```

READING AND UNDERSTANDING THE DATASET

jupyter LeadScoringCaseStudy Last Checkpoint: 11 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

2. Reading and undertanding the Dataset

In [3]:
```
leaddata = pd.read_csv("Leads.csv")
```

In [4]:
```
leaddata.head()
```

Out[4]:

	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	C
0	7927b2df-8bba-4d29-b9a2-	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Sel

DATA PREPROCESSING & EDA

Run Markdown

```
dtype: float64
```

In [10]:
```
# considering 30% is the threshold of accepatble null values in a column
#if the null values in a column exceeds approx 30% then the column may not be useful for Analysis and
# henceforth we are going to drop those columns from Analysis
# we are considering actual number rather than % because we have around 9K records
colswith2800nulls= (leaddata.columns[leaddata.isnull().sum() >= 2800]).tolist()
print('Number of Columns which is eligible t be dropped as per Assumption :',len(colswith2800nulls))
print('columns to be dropped ',colswith2800nulls)
```

```
Number of Columns which is eligible t be dropped as per Assumption : 6
columns to be dropped  ['Tags', 'Lead Quality', 'Asymmetrique Activity Index', 'Asymmetrique Profile
Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score']
```

In [11]:
```
# dropping columms
leaddatamod = leaddata.drop(colswith2800nulls,axis=1)
```

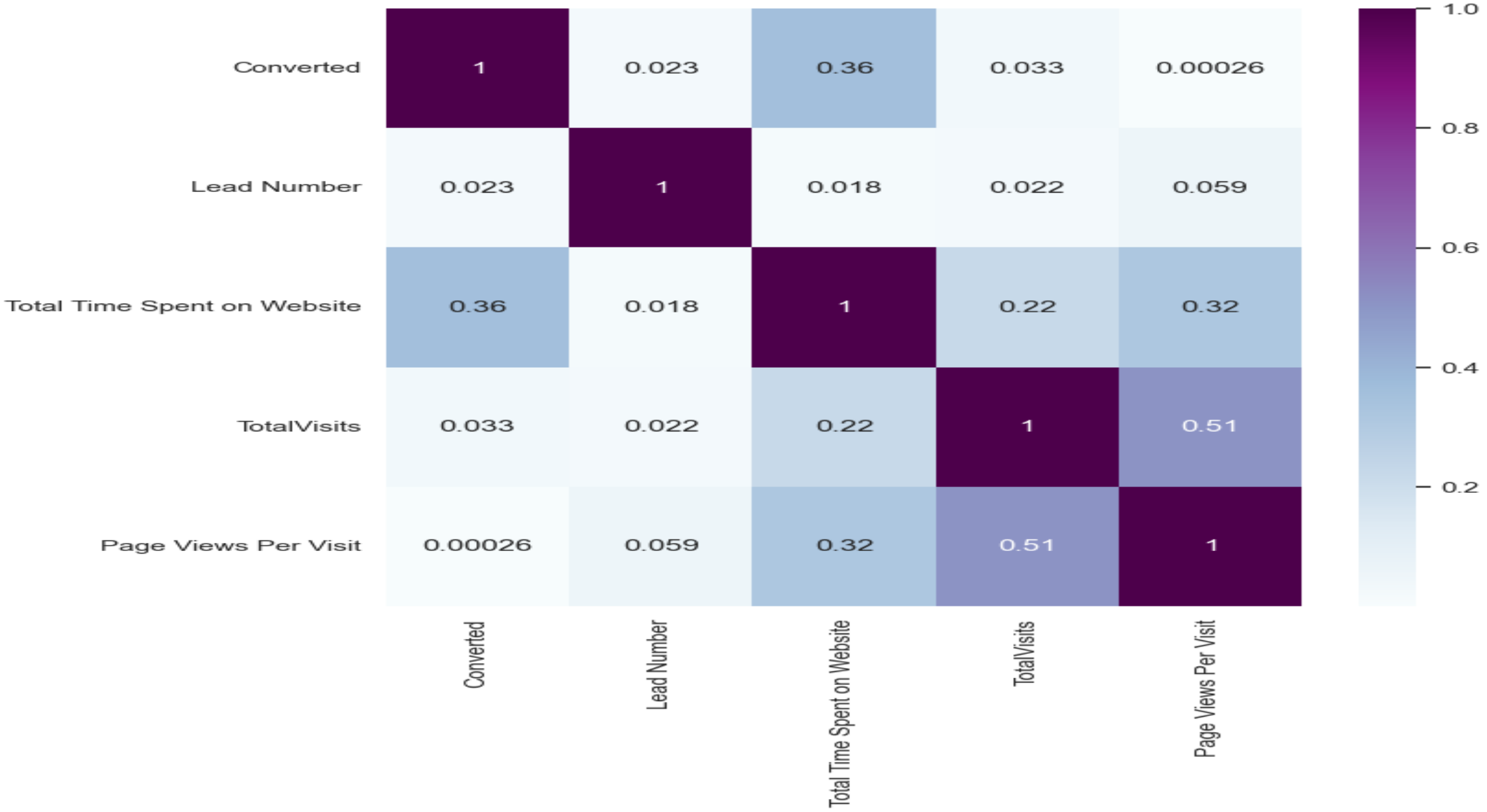
DATA PREPROCESSING & EDA

Null value handling with rationale

1. Analysis of each column having null vakues and their plan of handling
2. Lead Source >> Important attribute and if the Source is not known we can drop that records
3. TotalVisits >> Important attribute to score lead and the value is null then we can drop that records
4. Page Views Per Visit >> Important attribute to score lead and the value is null then we can drop that records
5. Last Activity >> Activity last done with the education organization, Important attribute to score lead and the value is null then we can drop that records
6. Country >> Since the company sells courses online; country has around 2.5K nulls and 6K just india henceforth this attribute may not add value and we can drop
7. Specialization >> total columns 9K , Null > 1438 and value select = 1942 i.e allmost 3.5K recs have no value henceforth we can drop
8. How did you hear about X Education >> total columns 9K , Null > 2204 and value select = 5043 i.e allmost 7K records have no value henceforth we can drop this
9. What is your current occupation >> Occupation could be factor to choose courses , around 2690 is NULL, we will impute with Mode
10. What matters most to you in choosing >> total columns 9K , Null > 2709 and Value Better Career Prospects = 6528 and no other prominents values , seems will not impact decesion and we can drop
11. Lead Profile >> total columns 9K , Null > 2709 and value select = 4146 i.e allmost 7K records have no value henceforth we can drop this
12. City >> total columns 9K , Null > 1420 and value select = 2249 i.e allmost 4K recs have no value henceforth we can drop

DATA PREPROCESSING & EDA

Correlation Heatmap for continuous variables



DATA PREPROCESSING & EDA

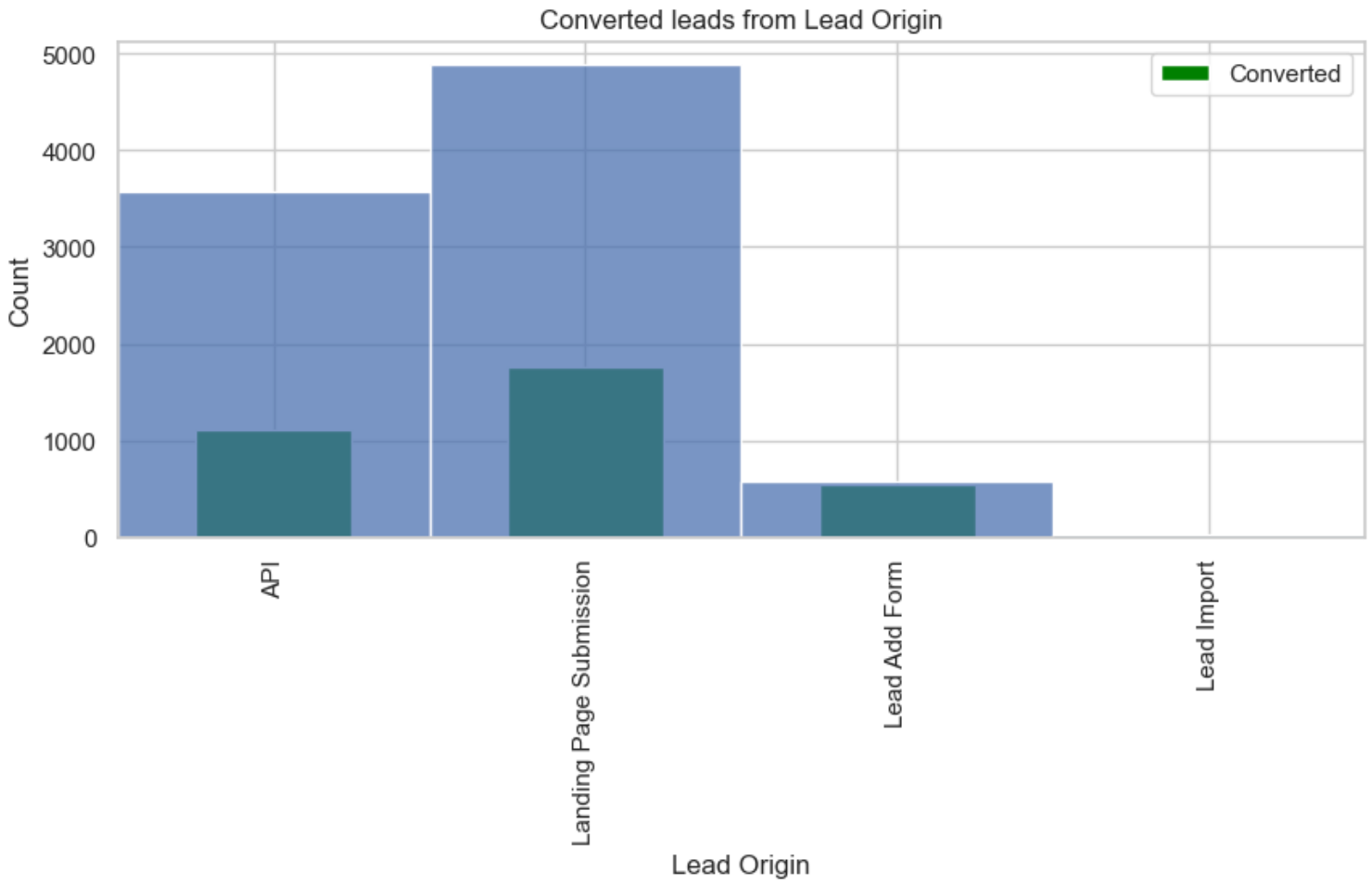
Attributes which do not have proper value to help the analysis and henceforth do not add any value

Analysis based on above visuals (refer Python file) and values printed below has to be done

1. we can drop the below columns as only one value exists and that will not have any impact on analysis or Decesion making
2. Do Not Call (no = 9072 , yes = 2)
3. Search (no = 9060 , yes = 14)
4. Magazine (no=9074)
5. Newspaper Article (no = 9072 , yes = 2)
6. X Education Forums (no = 9073 , yes = 1)
7. Newspaper (no = 9073 , yes = 1)
8. Digital Advertisement (no = 9070 , yes = 4)
9. Through Recommendations (no = 9067 , yes = 7)
10. Receive More Updates About Our Courses (no=9070)
11. Update me on Supply Chain Content (no = 9074)
12. Get updates on DM Content (no=9074)
13. I agree to pay the amount through cheque (no=9074)

DATA PREPROCESSING & EDA

Lead Origin with respect to conversions



MODEL BUILDING & EVALUATION

Creating Dummies

jupyter LeadScoringCaseStudy Last Checkpoint: 10 minutes ago (autosaved)

geeksforgeeks.org
This tab is sleeping to save resources.
Estimated savings: 85%

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

In [56]:
```
#Making Categorical to by dumpfication
#makedumycols = ['Lead Origin']
# checking if Drop first is reuired or not, with drop firt we are loosing the information
leaddatamod_logreg = pd.get_dummies(data=leaddatamod,columns=makedumycols)
```

In [57]:
```
leaddatamod_logreg.columns
```

Out[57]:
```
Index(['Converted', 'TotalVisits', 'Total Time Spent on Website',
       'Page Views Per Visit', 'Lead Origin_API',
       'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form',
       'Lead Origin_Lead Import', 'Lead Source_Click2call',
       'Lead Source_Direct Traffic', 'Lead Source_Facebook',
       'Lead Source_Google', 'Lead Source_Live Chat', 'Lead Source_NC_EDM',
       'Lead Source_Olark Chat', 'Lead Source_Organic Search',
       'Lead Source_Pay per Click Ads', 'Lead Source_Press_Release',
       'Lead Source_Reference', 'Lead Source_Referral Sites',
       'Lead Source_Social Media', 'Lead Source_WeLearn',
       'Lead Source_Welingak Website', 'Lead Source_bing', 'Lead Source_blog',
```

MODEL BUILDING & EVALUATION

Scaling of Numeric/ Continuous Variables

jupyter LeadScoringCaseStudy Last Checkpoint: 11 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

Scaling the Numeric / continous Variable

In [62]:
```
# scaling of numerical variables for Train data
scale = MinMaxScaler()
numcols1 = ['Total Time Spent on Website', 'TotalVisits', 'Page Views Per Visit']
X[numcols1] = scale.fit_transform(X[numcols1])
```

In [63]:
```
## spliting between train and test
# train abd test split ( 70:30  )
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=.3,random_state=41)
```

In [64]:
```
# checking on the shapes of the train and test dataset
X_train.shape,X_test.shape,y_train.shape,y_test.shape
```

Out[64]:
```
((6351, 71), (2723, 71), (6351,), (2723,))
```

MODEL BUILDING & EVALUATION

Creating Logistic Regression Model using Sklearn

jupyter LeadScoringCaseStudy Last Checkpoint: 12 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

Creating a Base Model for LR using Sklearn (Model : 1)

In [65]:
```
#building a model using Sklearn sort of base model
logreg = LogisticRegression(random_state=41)
```

In [66]:
```
logreg.fit(X_train, y_train)
```

Out[66]:
```
LogisticRegression
LogisticRegression(random_state=41)
```

In [67]:
```
y_train_pred = logreg.predict(X_train)
```

In [68]:
```
accuracy_score(y_train, y_train_pred)
```

Out[68]: 0.81561052746654[illegible]

MODEL BUILDING & EVALUATION

Model Evaluation using Cross Validation

jupyter LeadScoringCaseStudy Last Checkpoint: 12 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

Model Evaluation using Cross validation

In [73]:
```
cross_val_score(logreg, X_train, y_train, cv=5, n_jobs=-1)
```

Out[73]:
```
array([0.83005507, 0.82204724, 0.78740157, 0.81732283, 0.8023622 ])
```

In [74]:
```
cross_val_score(logreg, X_train, y_train, cv=5, n_jobs=-1).mean()
```

Out[74]:
```
0.8118377866024025
```

In [75]:
```
cross_val_score(logreg, X_test, y_test, cv=5, n_jobs=-1).mean()
```

Out[75]:
```
0.8145460064759849
```

MODEL BUILDING & EVALUATION

Feature selection Using RFE and Few models Created by varying Number of Feature Value

Feature selection & Creating few more models to compare (Model : 2)

Recursive Feature Elimination - RFE

In [76]:
```
logreg = LogisticRegression(random_state=41)
```

In [77]:
```
rfe = RFE(estimator=logreg, n_features_to_select=11)
```

In [78]:
```
rfe.fit(X_train, y_train)
```

Out[78]:
```
▸ RFE
▸ estimator: LogisticRegression
  ▸ LogisticRegression
```

MODEL BUILDING & EVALUATION

Model Evaluation is dne using Cross Validation Score for both Train and Test(unseen) set

jupyter LeadScoringCaseStudy Last Checkpoint: 15 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

```
X_train3.shape
```

Out[90]: (6351, 15)

In [91]:
```
cross_val_score(logreg, X_train3, y_train, n_jobs=-1).mean()
```

Out[91]: 0.7951482185891201

In [92]:
```
# Testing in test Data / Unseen data
X_test3 = X_test.loc[:,rfe.support_]
X_test3.shape
```

Out[92]: (2723, 15)

In [93]:
```
cross_val_score(logreg, X_test3, y_test, n_jobs=-1).mean()
```

Out[93]: 0.7914145979492714

MODEL BUILDING & EVALUATION

Model Evaluation is done using Cross Validation Score for both Train and Test(unseen) set

jupyter LeadScoringCaseStudy Last Checkpoint: 15 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

```
X_train3.shape
```

Out[90]: (6351, 15)

In [91]:
```
cross_val_score(logreg, X_train3, y_train, n_jobs=-1).mean()
```

Out[91]: 0.7951482185891201

In [92]:
```
# Testing in test Data / Unseen data
X_test3 = X_test.loc[:,rfe.support_]
X_test3.shape
```

Out[92]: (2723, 15)

In [93]:
```
cross_val_score(logreg, X_test3, y_test, n_jobs=-1).mean()
```

Out[93]: 0.7914145979492714

CONCLUSION / SUMMARY

Summary of the Model buding excercise
Conclusion

below are he cross val score for the few model we had build and analysed

Model(s)	Cross Val Score(Train)	cross Val score (Test)
Base Model	0.8118377866024025	0.8145460064759849
Model 2	0.8003434582478922	0.7983917970858069
Model 3	0.7951482185891201	0.7914145979492714
Model 4	0.7975099277027822	0.7950876956287102
Model 5	0.8126255598852661	0.8134504856988667
Model 6	0.8121531189403843	0.813815434430653
Model 7	0.8009733795077347	0.7987580949811118

Model 5 looks to be best model below are the features of the model 5 , 25 features considered below seems to be optimal [*'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'Lead Origin_API', 'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Reference', 'Lead Source_ Welingak Website', 'Do Not Email_No', 'Last Activity_Approached upfront', 'Last Activity_Converted to Lead', 'Last Activity_Email Bounced', 'Last Activity_Had a Phone Conversation', 'Last Activity_Olark Chat Conversation', 'What is your current occupation_Housewife', 'What is your current occupation_Student', 'What is your current occupation_Unemployed', 'What is your current occupation_ Working Professional', 'Last Notable Activity_Email Link Clicked', 'Last Notable Activity_Email Opened', 'Last Notable Activity_Had a Phone Conversation', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversation', 'Last Notable Activity_Page Visited on Website'*]

CONCLUSION / SUMMARY

1. Lead Origin and Lead source are key for getting the lead converted
2. Leads Originating from the below source
   2.1 API
   2.2 Landing Page Submission
   2.3 Add form
3. Lead Sources coming from below
   3.1 Olark chat
   3.2 Reference
   3.3 website
4. Company should consider to reach out leads from point 2 and 3 as they are potential leads to be converted
5. Below leads whose occupation are considered to be potential leads to be converted
   5.1 unemployed
   5.2 Student
   5.3 house wife
   5.4 working professional

THANK YOU