Lead Scoring Case Study

Aswin Kumar. K

Alankita Kumari

Amarnath Marelle

Problem Statement

- X Education – an online education company is facing problems to convert the leads generated.
- Typical conversion rate of X Education leads is currently 30%
- X Education wants the consultant to build a model to take the target lead conversion rate to around 80%.

Methodology used for the study

01. Data cleaning and preparation

Handling missing values

Mapping categorical variables to integers

Dummy variables creation

02. Test-train split and Scaling

Data split into 80 to 20 ratio

Usage of standard scaler

03. Model Building

Feature selection using RFE

Manual feature elimination (using p-values and VIFs)

04. Model Evaluation

Sensitivity and Specificity

Optimal cutoff using ROC

Precision and Recall

05. Prediction on test set

Final model testing for results

01. Data cleaning and preparation

Total data – 37 columns and 9240 rows

30 columns/categories are object type

Missing values columns dropping

Columns with high missing values and low significance (data label indexes) are dropped

categorical variables have a level called 'Select' handled

Three columns dropped for having significant number of data points as ‘Select’

Columns carrying in insignificant data dropped

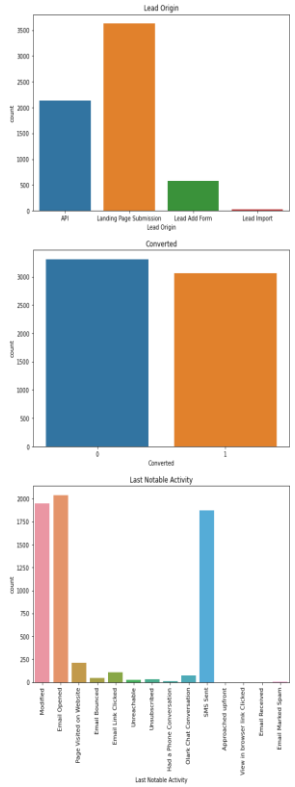
Columns having ‘no’ as maximum results (more than 95%) and majorly one result are dropped from dataset

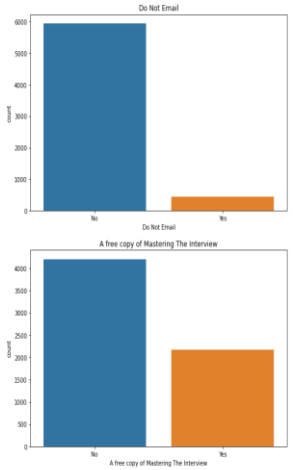
Handling the null values

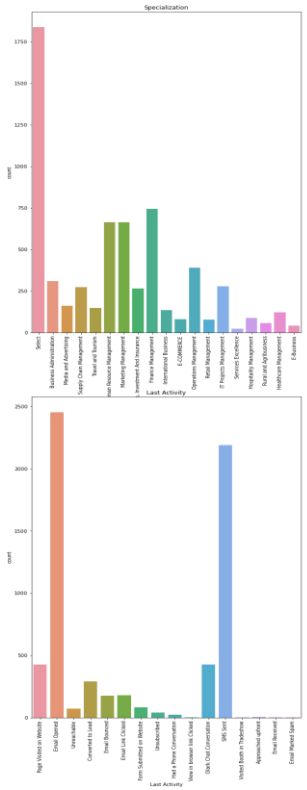
Replacing the null values in columns and dropping the columns/rows as needed

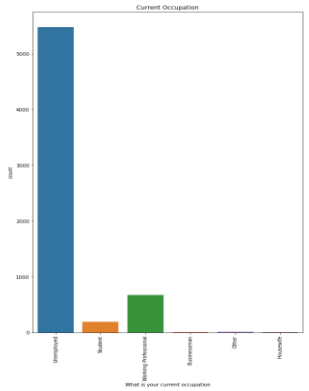
01. Data cleaning and preparation

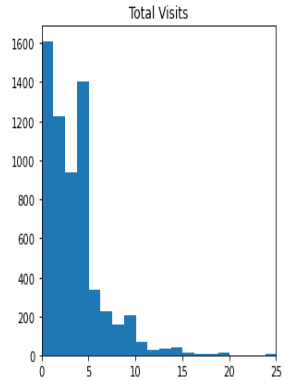
Exploratory Data analysis data visualizations

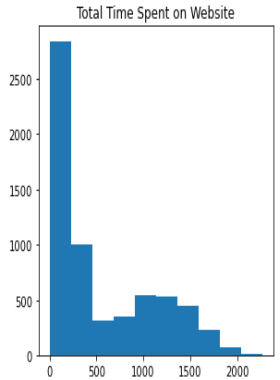


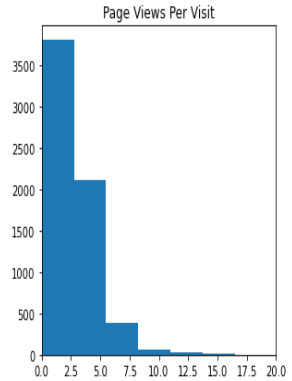


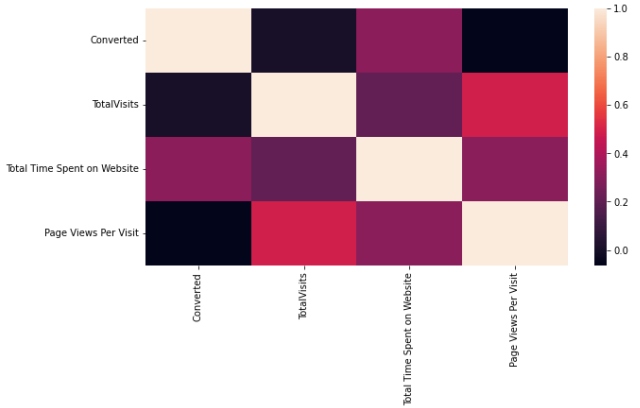












01. Data cleaning and preparation

Dummy variables creation

8 categorical variables in the data frame – Dummy variables created for each category

Separate dummy variable for ‘Specialization’ column

Separating the value 'select' from the variable 'Specialization' and creating a separate dummy variable

Dummy variable creation

There are 8 categorical variables in the data frame. Dummy variables are to be created for these categories

```
[ ] dummy = pd.get_dummies(leads[['Lead Origin', 'Lead Source', 'Do Not Email',
          'Last Activity', 'What is your current occupation',
          'A free copy of Mastering The Interview', 'Last Notable Activity']], drop_first=True)
```

```
[ ] leads = pd.concat([leads, dummy], axis=1)
```

```
[ ] ## Separating the value 'select' from the variable 'Specialization' and creating a separate dummy variable

    dummy_spl = pd.get_dummies(leads['Specialization'], prefix = 'Specialization')
    dummy_spl = dummy_spl.drop(['Specialization_Select'], 1)
    leads = pd.concat([leads, dummy_spl], axis = 1)
```

```
[ ] ##Dropping the variables for which the dummy variables have been created

    leads = leads.drop(['Lead Origin', 'Lead Source', 'Do Not Email',
```

02. Test-train split and Scaling

Test-train split

Total data split in 80 to 20 ratio

i.e., train size: 0.80 and test size: 0.20

Data Scaling

Data processing using Standard Scaler done for three columns such as

TotalVisits

Total Time Spent on Website

Page Views Per Visit

```
[ ] from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()

num_cols=X_train.select_dtypes(include=['float64', 'int64']).columns

X_train[num_cols] = scaler.fit_transform(X_train[num_cols])

X_train.head()
```

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Live Chat	...	Specialization_ Project Managemer
4719	0.261194	1.601468	0.021340	1	0	0	0	0	1	0	...	
4453	-0.309804	1.848085	-0.222350	1	0	0	1	0	0	0	...	
5770	-0.690470	-0.946316	-1.197114	0	0	0	0	0	0	0	...	
1774	-0.119471	1.296302	0.265031	1	0	0	0	0	0	0	...	
403	-0.309804	1.303399	-0.222350	0	0	0	0	0	1	0	...	

03. Model Building

Feature selection using RFE

RFE was employed to select **15 variables for the model**

Manual feature elimination (using p-values and VIFs)

Total of 5 iterations using the elimination based on p-values and VIFs was used to arrive at the final model with **11 variables for the model**

```
[ ] # RFE selected columns grouped as one
    group = X_train.columns[rfe.support_]
    group
```

```
Index(['Total Time Spent on Website', 'Lead Origin_Lead Add Form',
       'Lead Source_Olark Chat', 'Lead Source_Reference',
       'Lead Source_Welingak Website', 'Do Not Email_Yes',
       'Last Activity_Had a Phone Conversation', 'Last Activity_SMS Sent',
       '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_Had a Phone Conversation',
       'Last Notable Activity_Modified', 'Last Notable Activity_Unreachable'],
      dtype='object')
```

04. Model Evaluation

Accuracy, Sensitivity and Specificity

Initial arbitrary 0.50 cut off was used to test the data

Accuracy of 79%; Sensitivity and specificity of 74% and 84% respectively is seen for model

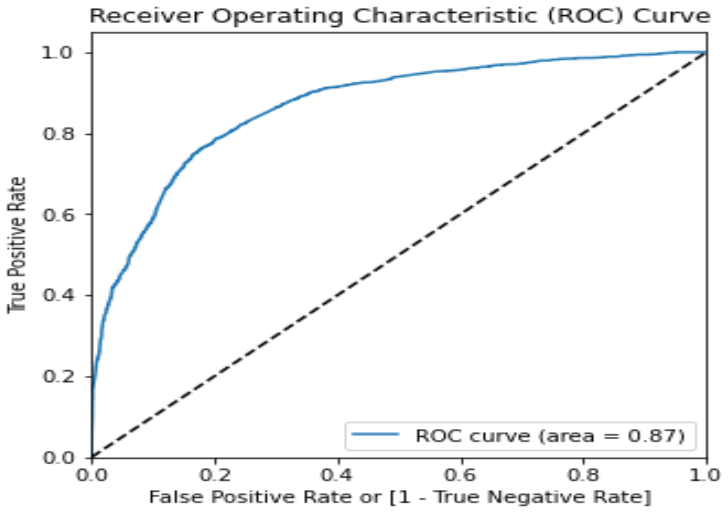
Precision and Recall

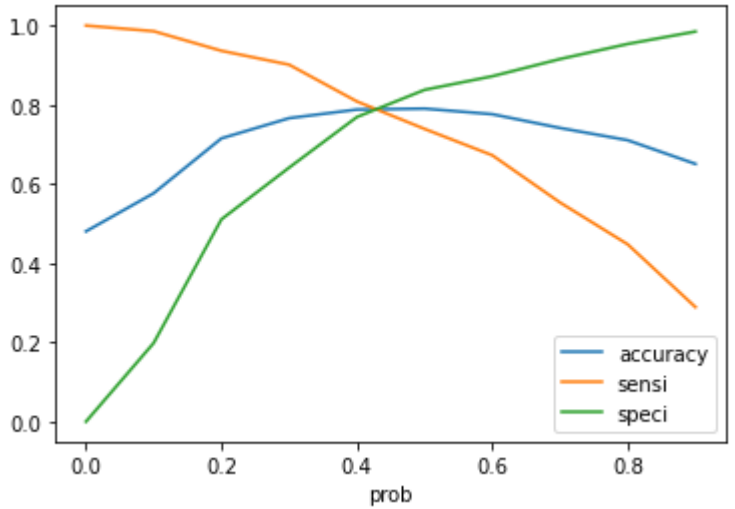
Accuracy of 79%; Precision and Recall of 78% and 79% respectively is seen for model

Optimal cutoff using ROC

ROC curve area = 0.87

Optimal cutoff point seen to be at 0.42





05. Prediction on test set

Predictions on test set based on the final model

Final conversion probability of 0.42 is set

Accuracy : 78%

Precision : 77%

Recall: 78%

Precision represents the number of correctly predicted positive instances and Recall represents model's ability to correctly identify all the positive instances in the dataset.

```
y_pred_final['final_predicted'] = y_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.42 else 0)
y_pred_final.head()
```

	Converted	Conversion_Prob	final_predicted
0	1	0.996498	1
1	0	0.148260	0
2	0	0.740229	1
3	1	0.388044	0
4	1	0.517452	1

```
[ ] metrics.accuracy_score(y_pred_final['Converted'], y_pred_final.final_predicted)
```

```
0.7803921568627451
```

Results Summary

Majorly 7 variables are seen having positive statistical correlations in predicting the hot leads

Lead Origin_Lead Add Form

Last Activity_Had a Phone Conversation

Last Notable Activity_Unreachable

Lead Source_Welingak Website

Lead Source_Olark Chat

Total Time Spent on Website

Last Activity_SMS Sent

However, Last Notable Activity_Unreachable can removed from key variables is it an unviable option despite the significance

Four variables can be used to categorize the leads into cold leads, which can help in focusing on the right leads and process efficient

Last Notable Activity_Modified

Do Not Email_Yes

What is your current occupation_Student

What is your current occupation_Unemployed

Thank you