Lead Scoring Case Study

- PRANIT SOMNE

Problem Statement

- Low Lead Conversion Rate: X Education currently experiences a typical lead conversion rate of only 30%, resulting in lost potential revenue and engagement.
- Inefficient Sales Outreach: The sales team spends considerable time contacting a broad range of leads, including those with low likelihood of conversion, which can dilute their effectiveness.
- Lack of Targeted Lead Identification: Without a systematic approach to identify high-potential leads (Hot Leads), the sales team cannot prioritize their efforts based on lead quality.
- Need for Data-Driven Strategy: To enhance conversion rates, X Education requires a data-driven strategy to analyze lead behaviors and characteristics, enabling more effective targeting and resource allocation.

Business Objective

- Objective of Lead Scoring Model: X Education aims to develop a model that assigns a lead score between 0 and 100 for each prospect, facilitating the identification of Hot Leads to enhance conversion rates.

- Target Conversion Rate: The CEO seeks to achieve an ambitious lead conversion rate of 80%, significantly higher than the current rate, to maximize revenue and engagement.

- Post-Target Strategies: Once the conversion target is achieved, the model should provide actionable insights and strategies for ongoing lead management and further improvement in conversion efforts.

Approach

- Data Cleaning and Preparation
- Model Building
- Model Evaluation
- Making Predictions on the Test Set
- Observations
- Conclusion

Data Cleaning and Preparation

- Null value handling – by eliminating few columns and by dropping missing value rows for few columns
- Dummy variable creation for categorical variables
- Train Test Split - Splitting the dataset into 70% train and 30% test
- Scaling numeric variables in the dataset with different scales to ensure consistency

The variable What matters most to you in choosing a course has the level Better Career Prospects appearing 6,528 times, while the other levels appear only once, twice, and once, respectively. Since this column is dominated by a single level and lacks variability, it is best to drop it.

```
In [38]: leads.drop('What matters most to you in choosing a course', axis=1, inplace=True)
```

```
In [39]: # Check the number of null values again
         leads.isnull().sum().sort_values(ascending=False)
```

```
Out[39]: What is your current occupation           2690
         Specialization                            1438
         TotalVisits                                137
         Page Views Per Visit                       137
         Last Activity                              103
         Lead Source                                 36
         Prospect ID                                  0
         Lead Number                                  0
         Lead Origin                                  0
         Do Not Email                                 0
         Converted                                    0
         Total Time Spent on Website                  0
         A free copy of Mastering The Interview       0
         Last Notable Activity                        0
         dtype: int64
```

Given that we have already removed many feature variables and What is your current occupation might still be significant, we will avoid dropping the entire column. Instead, we will only remove the rows with null values in the What is your current occupation column.

```
In [40]: # Drop the null value rows in the column 'What is your current occupation'
         leads = leads[~pd.isnull(leads['What is your current occupation'])]
```

```
In [62]: # We will create dummy variables for the Specialization column separately.
         # Since the level Select is not useful, we will explicitly drop that level before generating the dummy variables.
         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)
```

```
In [67]: # Spliting the dataset into 70% train and 30% test
         X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100)
```

```
In [50]: # Scaling the three numeric features present in the dataset
         scaler = MinMaxScaler()
         X_train[['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']] = scaler.fit_transform(X_train[['TotalVisits', '
         X_train.head()
```

Model Building

- Given the large number of variables, we can effectively reduce the feature set by using Recursive Feature Elimination (RFE). This method will help us select a smaller and more relevant set of features from the pool of variables for our model.
- Iteratively building models by considering VIF and p-values

Out[90]:

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

Out[89]:

Generalized Linear Model Regression Results

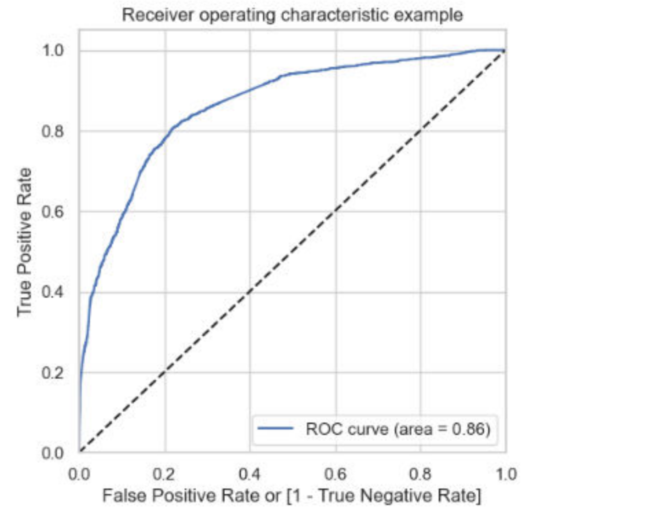
Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4449
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2079.1
Date:	Sat, 21 Sep 2024	Deviance:	4158.1
Time:	15:28:08	Pearson chi2:	4.80e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3642
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

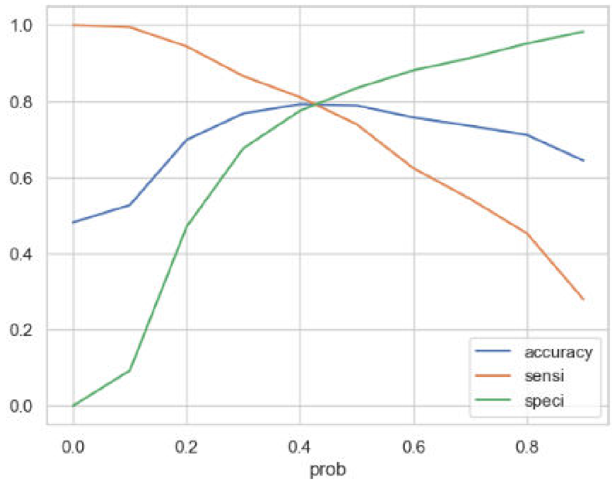
Model Evaluation

- ROC Curve - The area under the curve of the ROC is 0.86 which is quite good.
- Cut-off between accuracy, sensitivity and specificity come to 0.42
- After calculating precision and recall new cut-off comes to 0.44

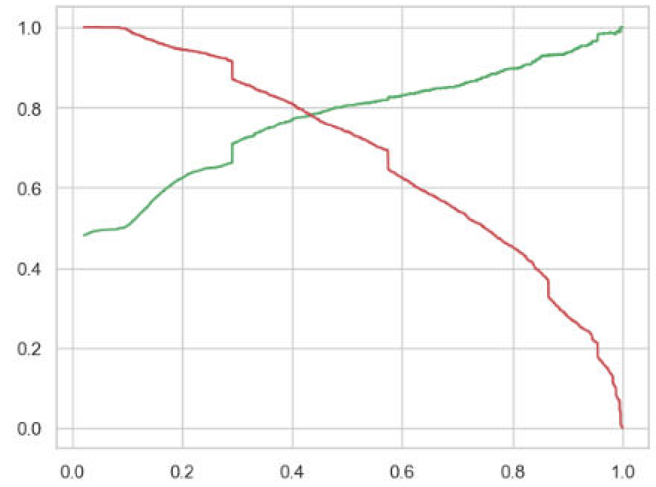
In [102]:
```
# Calling the ROC function
draw_roc(y_train_pred_final.Converted, y_train_pred_final.Conversion_Prob)
```



In [105]:
```
# Let's plot it as well
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```



In [139]:
```
plt.plot(thresholds, p[:-1], "g-")
plt.plot(thresholds, r[:-1], "r-")
plt.show()
```



Making Predictions on the Test Set

- Made predictions on test data using final model
- Used 0.44 as the cut-off on the predicted values
- Precision comes to 0.78
- Recall comes to 0.76

```
In [161]: # Calculating the Precision

          TP/(TP+FP)

Out[161]: 0.7828507795100222

In [162]: # Calculating Recall

          TP/(TP+FN)

Out[162]: 0.767467248908297
```

Observations

Train Data:

Accuracy : 78.45%

Sensitivity : 77.95%

Specificity : 78.92%

Test Data:

Accuracy : 78.66%

Sensitivity : 76.75%

Specificity : 80.42%

Final Feature List:

- What is your current occupation_Unemployed
- Total Time Spent on Website
- TotalVisits
- Last Activity_SMS Sent
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Lead Source_Welingak Website
- Do Not Email_Yes
- What is your current occupation_Student
- Last Activity_Had a Phone Conversation
- Last Notable Activity_Unreachable

Conclusion

- **<u>Lead Conversion Focus:</u>** The logistic regression model was used to identify high-potential leads for X Education, aiming to improve conversion rates.

- **<u>Data Quality Enhancement:</u>** Addressed the issue of missing 'Select' entries in key fields like occupation and specialization.

- **<u>Engagement Correlation:</u>** Metrics such as total visits and time spent on the platform were strong indicators of lead conversion.

- **<u>Targeted Outreach:</u>** Marketing efforts should focus on specializations like HR, Finance, and Marketing, and on enhancing engagement via email and SMS for higher conversion rates

Thank You