Leads_Scoring_Case_Study

- Bawana Tikoo, Bharat Bhushan and Shilpa Gupta

Problem Statement

- The company requires a model wherein we 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 target lead conversion rate to be around 80%.
- We will build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is most likely to convert whereas a lower score would mean that the lead mostly will not get converted.

Details of the Data Set

This dataset has 2 files as explained below:

- ‘Leads.csv’ contains all the information of the leads dataset with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted
- 'Leads Data Dictionary.csv' contains information about the dataset Leads.csv and also the explanation of all columns.

Analysis Approach

- We will proceed with data cleaning first in order to identify the columns which can actually help us to gain insights in to the model building solution.
- We will include the specific columns which can help to get to the decision making in the case of lead conversion.
- We will perform dummification of the columns wherever feasible and include in model building tactically.
- Using logistic regression we will build the model, using RFE we will identify the columns which can help build a model which could further help us in in making a decision system for lead conversion.
- To identify the columns is the most important task from business perspective as well, as this will help to achieve the target lead conversion which is expected.
- We will consider p-values and VIFs as the important metrices to build the best model.
- We will be using specificity, sensitivity and accuracy to showcase how the derived model is the best fit and can do really well in making predictions.
- We will also provide with ROC graph to verify our model as the best fit model.

Data Cleaning (Leads.csv)

- We have imported the necessary libraries and read the data from the Leads.csv data set
- The data set comprises of rows and columns as (9240,37)
- We can check what columns are present and check for description of the dataset.
- As we see the number of rows and columns are quite high , we will go further and check for any missing values in the data set.

Missing values Treatment

- As , we have found a lot of missing values in the columns ,the maximum missing values found in a column are more than 4000.
- Dropping columns with more than 3000 missing values, as imputing large amounts of missing data could lead to bias or skewness.
- As ‘city’ & ‘Country’ has more than 25% of missing values & City is not an important feature for our analysis so we can drop it.
- There are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are missing values.

Treating ‘Select’ value in columns

- We will get ‘value counts’ of each column to determine the ‘Select’ value count for the columns as this is as good as missing values.
- 'Lead Profile' and 'How did you hear about X Education' have a lot of rows which have the value Select which is of no use to the analysis so we will drop them.

```
data['Lead Profile'].value_counts()
```

```
Lead Profile
Select                         4146
Potential Lead                 1613
```

```
data['How did you hear about X Education'].value_counts()
```

```
How did you hear about X Education
Select                   5043
Online Search             808
Word Of Mouth             348
```

Removing columns with no value diversity

- There are a few columns in which only one value was majorly present for all the data points. These include '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'. All of the values for these variables are 'No', so, it is best that we drop these columns as they won't be helpful with our analysis.

Dummification of Categorical Columns

Creating dummy variable separately for the variable 'Specialization' since it has the level 'Select' ,which is useless so drop that level by specifying it explicitly.

```
dummy_spl = pd.get_dummies(data['Specialization'], dtype=int)
dummy_spl = dummy_spl.drop(['Select'], axis=1)
data = pd.concat([data, dummy_spl], axis = 1)
```

Model Building

- Test-Train Split: Put all the feature variables in X and Put the target variable('Converted') in y.
- Split the dataset into 70% train and 30% test, and set the random state to 100.
- Scale the three numeric features: 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'.
- There are a lot of variables present in the dataset. So, the best way to approach this is to select a small set of features from this pool of variables using RFE.
- Using RFE and we are selecting 15 variables here to build the model using logistic regression.
- Fit a logistic Regression model on X_train after adding a constant and output the summary

Summary Output

Model 1 : Summary results

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	7810
Model:	GLM	**Df Residuals:**	7794
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3491.1
Date:	Tue, 22 Oct 2024	**Deviance:**	6982.2
Time:	03:20:08	**Pearson chi2:**	7.77e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3888
Covariance Type:	nonrobust		

p-value looks descent less than 0.05 for all variables

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9905	0.229	26.175	0.000	5.542	6.439
Lead Source_Direct Traffic	-3.3016	0.180	-18.386	0.000	-3.654	-2.950
Lead Source_Facebook	-3.3859	0.448	-7.562	0.000	-4.263	-2.508
Lead Source_Google	-2.8194	0.178	-15.880	0.000	-3.167	-2.471
Lead Source_Olark Chat	-2.7411	0.184	-14.934	0.000	-3.101	-2.381
Lead Source_Organic Search	-3.1391	0.189	-16.579	0.000	-3.510	-2.768
Lead Source_Referral Sites	-3.5135	0.341	-10.301	0.000	-4.182	-2.845
Lead Source_Welingak Website	1.6961	0.738	2.299	0.021	0.250	3.142
Do Not Email_Yes	-1.5121	0.160	-9.476	0.000	-1.825	-1.199
What is your current occupation_Student	-3.0554	0.250	-12.215	0.000	-3.546	-2.565
What is your current occupation_Unemployed	-2.6678	0.159	-16.770	0.000	-2.980	-2.356
What is your current occupation_Unknown	-4.0664	0.170	-23.898	0.000	-4.400	-3.733
Last Notable Activity_Email Link Clicked	-1.2010	0.238	-5.049	0.000	-1.667	-0.735
Last Notable Activity_Modified	-1.3861	0.068	-20.359	0.000	-1.520	-1.253
Last Notable Activity_Olark Chat Conversation	-1.7188	0.289	-5.944	0.000	-2.286	-1.152
Last Notable Activity_Page Visited on Website	-0.9681	0.177	-5.458	0.000	-1.316	-0.620

VIF observations

- VIFs seems to be in a decent range for all the variable as well.
- VIF & P-value seems descent enough for all the values where p-value is less than 0.05 & VIF is less than 5 proceed for Model Evaluation

	Features	VIF
9	What is your current occupation_Unemployed	4.02
10	What is your current occupation_Unknown	2.82
2	Lead Source_Google	2.53
0	Lead Source_Direct Traffic	2.33
3	Lead Source_Olark Chat	2.23
12	Last Notable Activity_Modified	1.66
4	Lead Source_Organic Search	1.59
6	Lead Source_Welingak Website	1.13
7	Do Not Email_Yes	1.13
8	What is your current occupation_Student	1.11
5	Lead Source_Referral Sites	1.07
13	Last Notable Activity_Olark Chat Conversation	1.07
14	Last Notable Activity_Page Visited on Website	1.06
11	Last Notable Activity_Email Link Clicked	1.03
1	Lead Source_Facebook	1.02

Model Evaluation

A dataframe with the actual conversion flag and the predicted probabilities:

A new dataframe containing the actual conversion flag and the probabilities predicted by the model.

	Converted	Conversion_Prob
0	0	0.201403
1	0	0.764373
2	0	0.309071
3	1	0.201403
4	1	0.203428

A new column 'Predicted' with 1 if Paid_Prob > 0.5 else 0

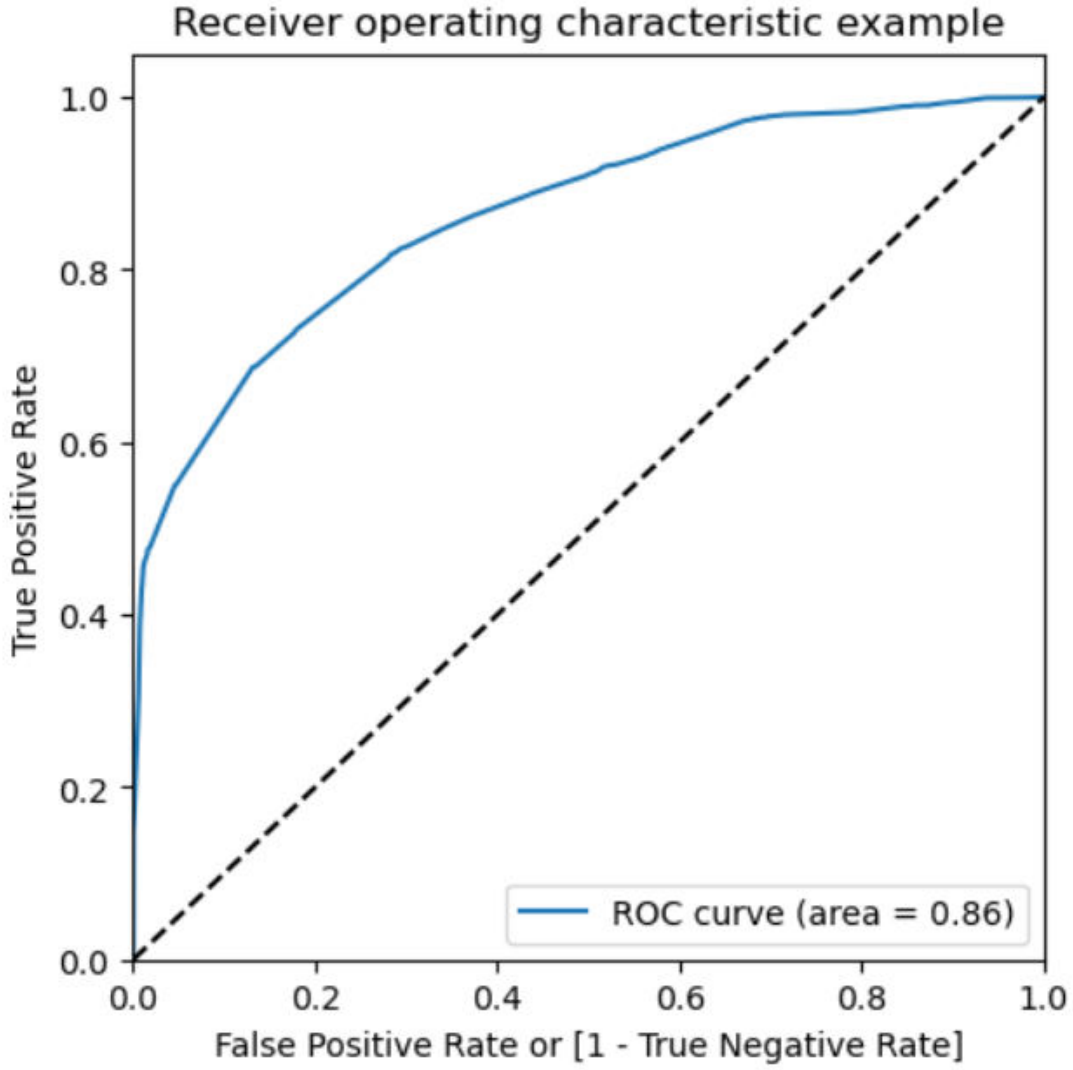
	Converted	Conversion_Prob	Predicted
0	0	0.201403	0
1	0	0.764373	1
2	0	0.309071	0
3	1	0.201403	0
4	1	0.203428	0

Confusion Matrix and Accuracy Score

- Accuracy score=76.65%
- Sensitivity=81.12%
- Specificity=72.18%
- False positive rate =27.81%
- Positive predictive value=74.47%
- Negative predictive value=79.27%
- With the current cut off as 0.5 we have around 76.6% accuracy with 81% sensitivity & 72% specificity. But in order to get good results, we need to optimise the threshold.

ROC

The area under the curve of the ROC is 0.86 is quite good. So, we seem to have a good model.



The sensitivity and specificity tradeoff to find the optimal cutoff point

Columns with different probability cutoffs

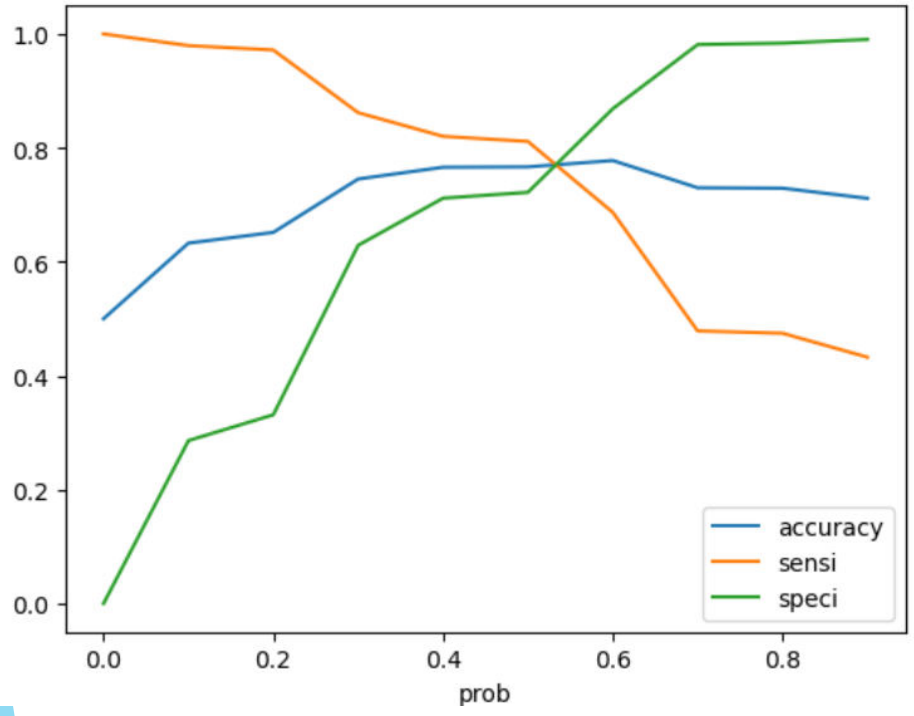
	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.201403	0	1	1	1	0	0	0	0	0	0	0
1	0	0.764373	1	1	1	1	1	1	1	1	1	0	0
2	0	0.309071	0	1	1	1	1	0	0	0	0	0	0
3	1	0.201403	0	1	1	1	0	0	0	0	0	0	0
4	1	0.203428	0	1	1	1	0	0	0	0	0	0	0

A dataframe to see the values of accuracy, sensitivity, and specificity at # different values of probability cutoffs

	prob	accuracy	sensi	speci
0.0	0.0	0.500000	1.000000	0.000000
0.1	0.1	0.632778	0.979513	0.286044
0.2	0.2	0.651601	0.971831	0.331370
0.3	0.3	0.745327	0.861716	0.628937
0.4	0.4	0.765941	0.820230	0.711652
0.5	0.5	0.766581	0.811268	0.721895
0.6	0.6	0.777721	0.686812	0.868630
0.7	0.7	0.729962	0.478617	0.981306
0.8	0.8	0.729193	0.474520	0.983867
0.9	0.9	0.711396	0.432522	0.990269

A plot to see the values of accuracy, sensitivity, and specificity at # different values of probabiity cutoffs

We can see that around 0.44, we get the optimal values of the three metrics. So let's choose 0.44 as our cutoff now. .



	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted
0	0	0.201403	0	1	1	1	0	0	0	0	0	0	0	0
1	0	0.764373	1	1	1	1	1	1	1	1	1	0	0	1
2	0	0.309071	0	1	1	1	1	0	0	0	0	0	0	0
3	1	0.201403	0	1	1	1	0	0	0	0	0	0	0	0
4	1	0.203428	0	1	1	1	0	0	0	0	0	0	0	0

Confusion Matrix and Accuracy Score on the optimal cut off

- Accuracy score=76.73%
- Sensitivity=81.12%
- Specificity=72.18%
- False positive rate =27.81%
- Positive predictive value=74.47%
- Negative predictive value=79.27%
- With the current cut off as 0.44, we have around 76.7% accuracy with 81% sensitivity & 72% specificity. The model looks good as well.

Predictions on the test set using 0.44 as the cutoff

	Converted	Conversion_Prob	final_predicted
0	0	0.289998	0
1	1	0.936368	1
2	0	0.289998	0
3	1	0.824768	1
4	0	0.228811	0

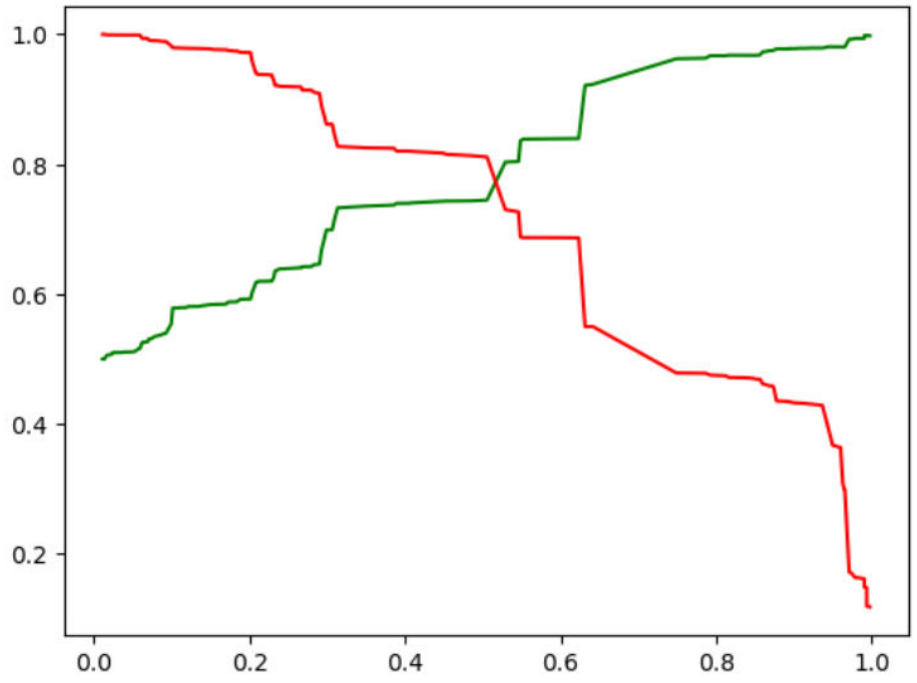
Confusion Matrix and Accuracy Score on the optimal cut off of .44 on test data.

- Accuracy score=71.94%
- Sensitivity=73.30%
- Specificity=71.16%
- Precision = 74.47%
- Recall = 81.12%

Precision & recall tradeoff

Confusion Matrix and Accuracy Score on the optimal cut off of .47

Accuracy score=76.64%
Precision = 74.36%
Recall = 81.33%
The higher recall value tell us that the model is more likely to predict the positive lead conversion.



	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	final_predicted
0	0	0.201403	0	1	1	1	0	0	0	0	0	0	0	0
1	0	0.764373	1	1	1	1	1	1	1	1	1	0	0	1
2	0	0.309071	0	1	1	1	1	0	0	0	0	0	0	0
3	1	0.201403	0	1	1	1	0	0	0	0	0	0	0	0
4	1	0.203428	0	1	1	1	0	0	0	0	0	0	0	0

Making final Predictions on the Test Set with .46 cut off

- Accuracy score=72.08%
- Precision = 59.45%
- Recall = 72.80%

	Converted	Conversion_Prob	final_predicted
0	0	0.289998	0
1	1	0.936368	1
2	0	0.289998	0
3	1	0.824768	1
4	0	0.228811	0

- The higher recall value tell us that the model is more likely to predict the positive lead conversion.

Key observations

- **Total Time Spent on Website**: Has the strongest positive impact on lead conversion. Leads that spend more time on the website are more likely to convert, highlighting the importance of website engagement.
- **What is your Current Occupation Working Professional**: Is a significant positive predictor. Working professionals tend to have a higher probability of converting into paying customers.
- **Lead Origin Lead Add Form**: Shows a strong positive influence. Leads generated through forms added to the website have a higher chances of conversion.

THANKYOU