LEAD SCORING
CASE STUDY
MASTER OF DATA SCIENCE, DS C44

Submitted by:

- **K Chandrahas, chandrahaskuridi007@gmail.com , 7090858334**
- **Thakur Yashraj Singh, yashrajsingh2301@gmail.com, 7207846570**
- **Jay Prakash George, jgjaygeorge@gmail.com , 9713471908**

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.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if 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.

Objective of the Case Study

- To 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 hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company in which the model should be able to adjust to if the company's requirement changes in the future. So it is needed to handle these as well.

Data Understanding and Preparation

- **Data Dimensions** : **(9240 rows, 37 columns)**
- **Data Imputation** :
  - ‘Select’ values were replaced with ‘Null’.
  - Values in some columns with ‘Null’ values were imputed with ‘Others’, ‘Not Sure’, etc. on a need-to-need basis.
  - Values in some columns which would result in a lot of unnecessary dummy variables were clubbed together and imputed with ‘Others’.
- **Null Value Handling** :
  - Columns with more than 70% null values were dropped.
- **Data Imbalance** :
  - Columns with High Data Imbalance were dropped from the analysis.

Data Preparation – Contd..

- **Dummy variables** were created for the categorical features.

- Data was divided into two parts for **Dependent & Independent variables** (on the basis of column '**Converted**').

  Data was split into **Train** & **Test** sets (random 70:30 split).

- Data Normalisation/Scaling was performed using **Min Max Scaler.**

Data Modelling

- **Feature Selection** :

  - RFE (Recursive Feature Elimination) : 15 variables were selected using RFE.

  - Manual Feature Selection : Logistic Regression was fitted on the Train data and p-values & VIF were calculated for each feature. Elimination happened recursively until we reached acceptable p-value & VIF for all the features.

Model Evaluation – Train data

- After training the model on Train set, predictions were made and their evaluation metrics were recorded.
- We kept a cut-off of 80% while making the predictions.
- Confusion Matrix was created :

Actual / Predicted	Not Converted	Converted
Not Converted	3879	74
Converted	490	1929

Data Preparation - Part 1

Data Imputation

- Replacing records with 'Select' values with Null.

In [252]:
```
leads.replace('Select', np.nan, inplace=True)
```

Checking for Null values

In [253]:
```
round((leads.isnull().sum()/len(leads.index))*100,2).sort_values(ascending=False)
```

Out[253]:

How did you hear about X Education	78.46
Lead Profile	74.19
Lead Quality	51.59
Asymmetrique Profile Score	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Profile Index	45.65
Asymmetrique Activity Index	45.65
City	39.71
Specialization	36.58
Tags	36.29
What matters most to you in choosing a course	29.32
What is your current occupation	29.11
Country	26.63
TotalVisits	1.48
Page Views Per Visit	1.48
Last Activity	1.11
Lead Source	0.39
Lead Origin	0.00
Lead Number	0.00
Do Not Email	0.00
Do Not Call	0.00

Data Preparation - Part 1

```
Receive More Updates About Our Courses            0.00
Update me on Supply Chain Content                 0.00
Get updates on DM Content                         0.00
I agree to pay the amount through cheque          0.00
Prospect ID                                       0.00
dtype: float64
```

Dropping columns with High percentage of Nulls

In [254]:
```
leads = leads.drop(['How did you hear about X Education','Lead Profile'],axis=1)
```

Checking column 'Lead Quality'

In [255]:
```
leads['Lead Quality'].value_counts(dropna=False)
```

Out[255]:
```
NaN                  4767
Might be             1560
Not Sure             1092
High in Relevance     637
Worst                 601
Low in Relevance      583
Name: Lead Quality, dtype: int64
```

Imputing Nulls with 'Not Sure'

In [256]:
```
leads['Lead Quality'].fillna('Not Sure',inplace=True)
```

In [257]:
```
plt.figure(figsize=(25,8))
sns.countplot(leads['Lead Quality'],order = leads['Lead Quality'].value_counts().index)
plt.show()
```

Data Preparation - Part 2

Creating Dummy Variables

In [299]:
```
dummies = pd.get_dummies(leads[['Lead Origin',
                                'Lead Source',
                                'Last Activity',
                                'Specialization',
                                'What is your current occupation',
                                'Last Notable Activity','Lead Quality','Tags','City',
                               'A free copy of Mastering The Interview']], drop_first=True)
```

In [300]:
```
#Concatenating dummy variables in the leads dataset
leads = pd.concat([leads, dummies], axis=1)
leads.head()
```

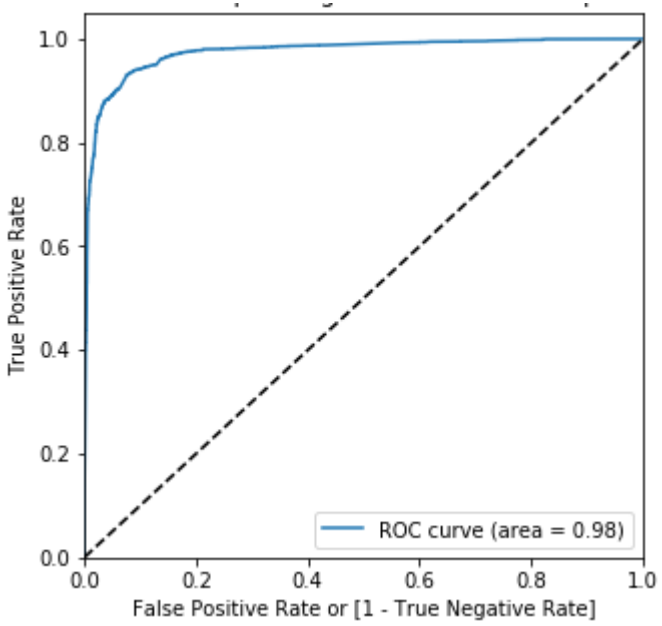
Out[300]:

	Lead Origin	Lead Source	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	A free copy of Mastering The Interview	Last Notable Activity	Orig
0	API	Olark Chat	0	0.0	0	0.0	Page Visited on Website	Other Specialization	Unemployed	Interested in other courses	Low in Relevance	Other Cities	No	Modified	
1	API	Organic Search	0	5.0	674	2.5	Email Opened	Other Specialization	Unemployed	Ringing	Not Sure	Other Cities	No	Email Opened	
2	Landing Page Submission	Direct Traffic	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be	Mumbai	Yes	Email Opened	
3	Landing Page Submission	Direct Traffic	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	No	Modified	
4	Landing Page Submission	Google	1	2.0	1428	1.0	Converted to Lead	Other Specialization	Unemployed	Will revert after reading the email	Might be	Mumbai	No	Modified	

Model Evaluation – Train data

- Evaluation Metrics :

Accuracy	Sensitivity	Specificity
91%	~80%	98%



ROC Curve

Observations :

The model seems to be performing well.
The ROC curve has a value of 0.98, which is very good.
Evaluation Metrics have a good balance as well.

Model Evaluation – Test data

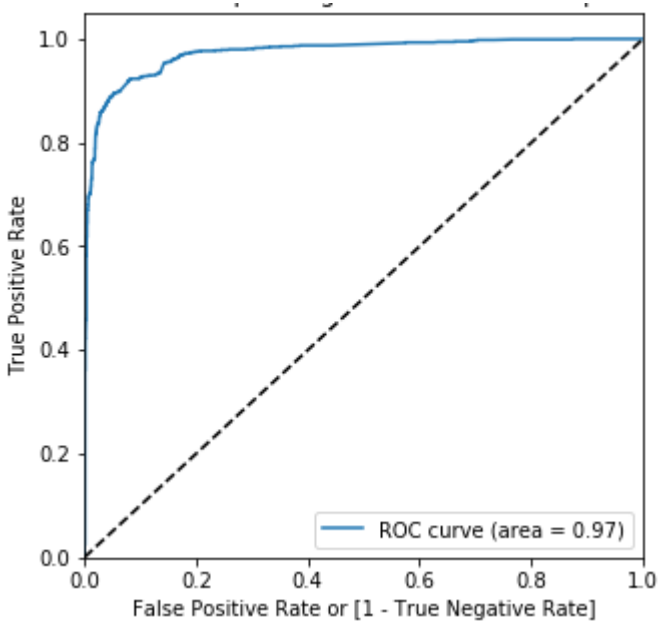
- After training the model on Train set, the predictions were made on the Test data set and their evaluation metrics were recorded.
- We kept a cut-off of 80% while making the predictions.
- Confusion Matrix was created :

Actual / Predicted	Not Converted	Converted
Not Converted	1656	33
Converted	215	827

Model evaluation – test data

- Evaluation Metrics:

Accuracy	Sensitivity	Specificity	Precision	Recall
~91%	79%	98%	96%	79%



ROC Curve

Observations :

The model seems to be performing well on test data as well.

The ROC curve has a value of 0.97, which is very good.

Evaluation Metrics have a good balance as well.

Conclusion & Recommendations

- The Model seems to be predicting the **Lead Conversation** very well, with a good balance of **all evaluation metrics** & good **ROC Curve** value.
- The Sales team can leverage this model with **high confidence** to call the leads.
- If the need arises for a model with **higher Sensitivity** (to make calls to only most potential leads), we can **increase** the **cut-off** even more.
- If the Sales team decides to **expand their Leads coverage**, **cut-off** can be **decreased** to meet the demands.