Lead Score Case Study

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

Data Dictionary:

Variables	Description
Prospect ID	A unique ID with which the customer is identified.
Lead Number	A lead number assigned to each lead procured.
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.
Converted	The target variable. Indicates whether a lead has been successfully converted or not.
TotalVisits	The total number of visits made by the customer on the website.
Total Time Spent on Website	The total time spent by the customer on the website.

Page Views Per Visit	**Average number of pages on the website viewed during the visits.**
Last Activity	**Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.**
Country	**The country of the customer.**
Specialization	**The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.**
How did you hear about X Education	**The source from which the customer heard about X Education.**
What is your current occupation	**Indicates whether the customer is a student, umemployed or employed.**
What matters most to you in choosing this course	**An option selected by the customer indicating what is their main motto behind doing this course.**
Search	**Indicating whether the customer had seen the ad in any of the listed items.**
Magazine	
Newspaper Article	
X Education Forums	
Newspaper	
Digital Advertisement	
Through Recommendations	**Indicates whether the customer came in through recommendations.**
Receive More Updates About Our Courses	**Indicates whether the customer chose to receive more updates about the courses.**
Tags	**Tags assigned to customers indicating the current status of the lead.**
Lead Quality	**Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.**

Update me on Supply Chain Content	**Indicates whether the customer wants updates on the Supply Chain Content.**
Get updates on DM Content	**Indicates whether the customer wants updates on the DM Content.**
Lead Profile	**A lead level assigned to each customer based on their profile.**
City	**The city of the customer.**
Asymmetrique Activity Index	**An index and score assigned to each customer based on their activity and their profile**
Asymmetrique Profile Index	
Asymmetrique Activity Score	
Asymmetrique Profile Score	
I agree to pay the amount through cheque	**Indicates whether the customer has agreed to pay the amount through cheque or not.**
a free copy of Mastering The Interview	**Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.**
Last Notable Activity	**The last notable activity performed by the student.**

Solution Approach:

After going through the data dictionary and understanding the problem at its core, a logistic regression model was decided fit for an optimal solution.

The solution was designed by following a series of steps involving but not limited to the following:

1. **Data Cleaning:**

The initial dataset was loaded and various quality and sanity checks were performed in order to bring the data in a format that would be a suitable fit for further analysis and modelling.

There were certain columns with a hefty amount of null values. The attributes with more than 40% null values were straight away dropped. Also, while giving a deeper look, it was observed that certain attributes had highly skewed data that would not have added any explanatory power to the model, they were dropped too. In some cases, a new category was created to account for missing data in order to prevent potential loss of information. Outliers were also removed.

After this process got completed, around 97% of data was retained.

2. **Exploratory Data Analysis:**

After getting our hands on a cleaner and more usable dataset, various plots were made to analyse and uncover the relationships different categorical and numerical attributes had with the target variable.

3. **Dummy Variables, Train - Test Split, and Feature Scaling:**

After the EDA, the phase of feature engineering kicked in. Here, dummy variables were created to account for categorical variable's explanatory power, and the data was then split in train and test sets.

Another important step was to scale the features so that all of them lie on a comparable scale.

4. **Model Building, Making Predictions and Model Evaluation:**

After the feature engineering part, the data was fit to be modelled.

First model: Logistic Regression:

The modelling part was done in 12 iterations where different models were built and evaluated for their significance. Checks of p-value significance and variance inflation factor(VIF) were done and a final version of the model was prepared.

After the model was built, predictions were made on it. In the evaluation part, 0.37 was decided as the optimal cut off point.

Metrics of the final model on test set:

i) Accuracy: 81%.
ii) Precision: 70%.
iii) Recall: 83%.

Second Model: Decision Tree Classifier

A Decision Tree Classifier model was built upon the parameters shortlisted by Recursive Feature Elimination earlier.

Metrics before Hyperparameter Tuning:

```
Train Accuracy : 0.9616009452075026
Train Confusion Matrix:
[[4063   69]
 [ 191 2448]]
--------------------------------------------------
Test Accuracy : 0.7936226749335695
Test Confusion Matrix:
[[1184  236]
 [ 230  608]]
```

Metrics After Hyperparameter Tuning:

```
Train Accuracy : 0.8243981686604638
Train Confusion Matrix:
[[3559  573]
 [ 616 2023]]
--------------------------------------------------
Test Accuracy : 0.8321523472099203
Test Confusion Matrix:
[[1228  192]
 [ 187  651]]
```

Feature Importances:

	Varname	Imp
2	Total Time Spent on Website	0.326766
3	LeadOrigin_Lead Add Form	0.303936
8	LastActivity_SMS Sent	0.131605
10	CurrentOccupation_No Information	0.088823
11	CurrentOccupation_Working Professional	0.085903
1	TotalVisits	0.039245
12	LastNotableActivity_Modified	0.018250
5	LeadSource_Welingak Website	0.003150
6	LastActivity_Email Opened	0.002323
0	Do Not Email	0.000000
4	LeadSource_Olark Chat	0.000000
7	LastActivity_Olark Chat Conversation	0.000000
9	LastActivity_Unsubscribed	0.000000
13	LastNotableActivity_Unreachable	0.000000

Third Model: Random Forest Classifier

Random Forest Classifiers are known for outperforming Decision Tree Classifiers because of the algorithm of using more than one Decision Tree to predict the outcomes.

Metrics before Hyperparameter Tuning:

```
Train Accuracy : 0.9616009452075026
Train Confusion Matrix:
[[4041   91]
 [ 169 2470]]
--------------------------------------------------
Test Accuracy : 0.8131089459698848
Test Confusion Matrix:
[[1209  211]
 [ 211  627]]
```

Metrics After Hyperparameter Tuning:

```
Train Accuracy : 0.838576281199232
Train Confusion Matrix:
[[3669  463]
 [ 630 2009]]
--------------------------------------------------
Test Accuracy : 0.8401240035429584
Test Confusion Matrix:
[[1249  171]
 [ 190  648]]
```

Feature Importance:

:

	Varname	Imp
2	Total Time Spent on Website	0.350828
10	CurrentOccupation_No Information	0.131998
3	LeadOrigin_Lead Add Form	0.127774
8	LastActivity_SMS Sent	0.119257
11	CurrentOccupation_Working Professional	0.090261
1	TotalVisits	0.057999
12	LastNotableActivity_Modified	0.046373
6	LastActivity_Email Opened	0.017270
7	LastActivity_Olark Chat Conversation	0.016635
4	LeadSource_Olark Chat	0.016261
0	Do Not Email	0.014372
5	LeadSource_Welingak Website	0.010671
9	LastActivity_Unsubscribed	0.000195
13	LastNotableActivity_Unreachable	0.000107