Lead Scoring Case Study

Submitted By :-

Name-Anish Kumar Thakur

Roll No-COE20018

Email id:- 20010594@cgu-odisha.ac.in

Agenda

The Purpose is to optimize the lead scoring mechanism based on their fit, demographic, behaviours , buying tendency etc. By implementing explicit and implicit scoring modelling with lead point system.

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.

The company markets its courses on several websites, search engines, and even social media sometimes. Once these people land on the website, they might browse the courses, fill out a form for the course, or watch some videos. When these people fill out a form with their email address or phone number, they are classified as leads. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted into successful sales, while most of the leads do not. The typical lead to successful sale conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead-to-sale conversion rate is very poor. For example, if they acquire 100 leads in a day, only about 30 of them are converted into successful sales. 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 would go up as the sales team would now be focusing more on communicating with the potential leads rather than making calls to everyone

Goals of the Case Study

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, Le. 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 which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well,

Approach

- Source the data For analysis
- Reading & Understanding the data
- Data Cleaning
- EDA
- Feature scaling
- Splitting the data into test & train dataset
- Prepare the data for modelling
- Model building
- Model evaluation-specificity & sensitivity or precision recall
- Making predictions on the test set

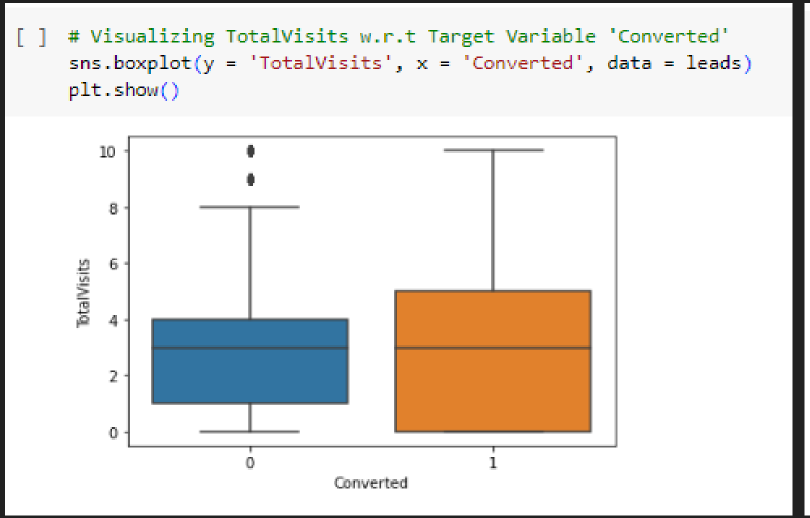
Data Sourcing ,Cleaning and Presentation

- Read the data from CSV File
- Outlier treatment
- Data cleaning Handling Null Values & removing higher Null values data
- Removing Redundant columns in the data
- Imputing Null Values
- Exploratory data analysis-approx. Conversion Rate is 38%
- Feature standardization

Outlier

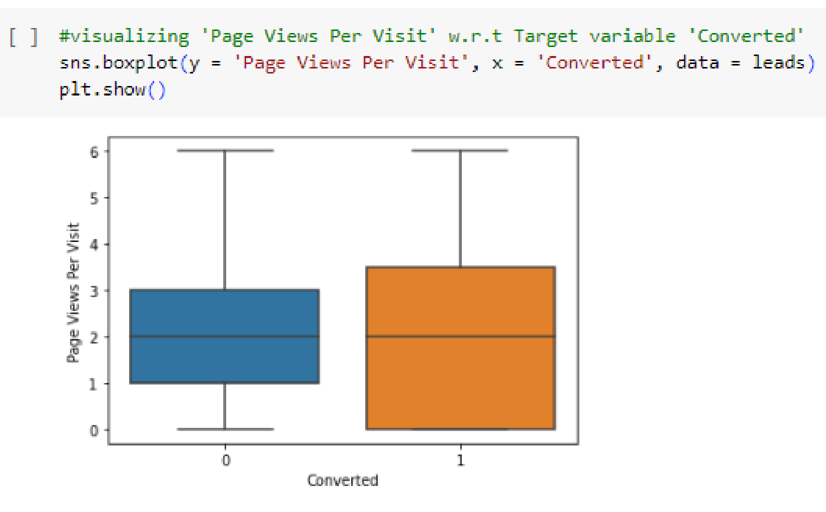
Total Visits, Total Time Spent on Website, Pages view per visits have outlier

Total Visits



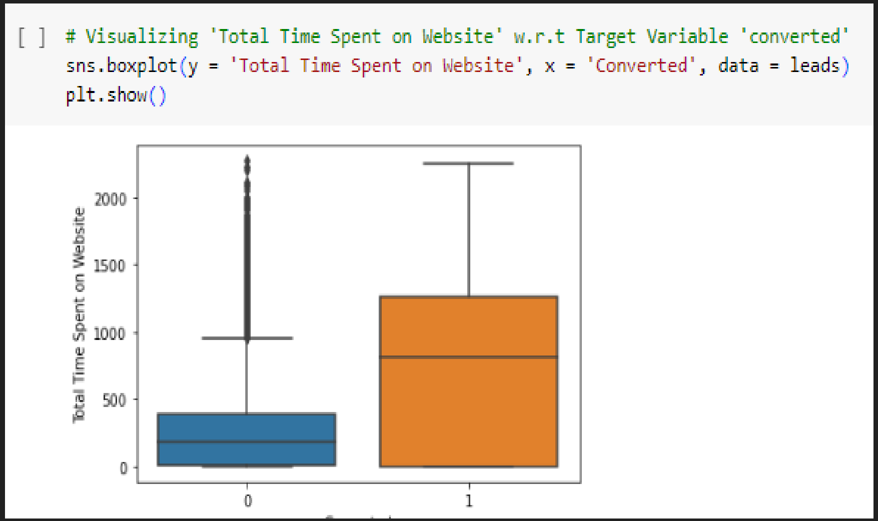
```
[ ]  # Visualizing TotalVisits w.r.t Target Variable 'Converted'
     sns.boxplot(y = 'TotalVisits', x = 'Converted', data = leads)
     plt.show()
```

Total Time Spent on Website



```
[ ]  #visualizing 'Page Views Per Visit' w.r.t Target variable 'Converted'
     sns.boxplot(y = 'Page Views Per Visit', x = 'Converted', data = leads)
     plt.show()
```

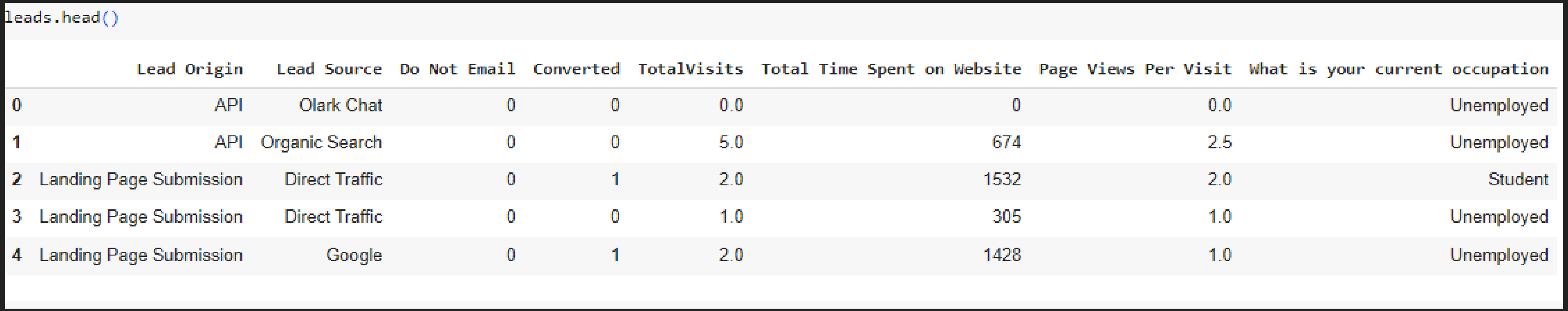
Pages view per visits



```
[ ]  # Visualizing 'Total Time Spent on Website' w.r.t Target Variable 'converted'
     sns.boxplot(y = 'Total Time Spent on Website', x = 'Converted', data = leads)
     plt.show()
```

Data Preparation

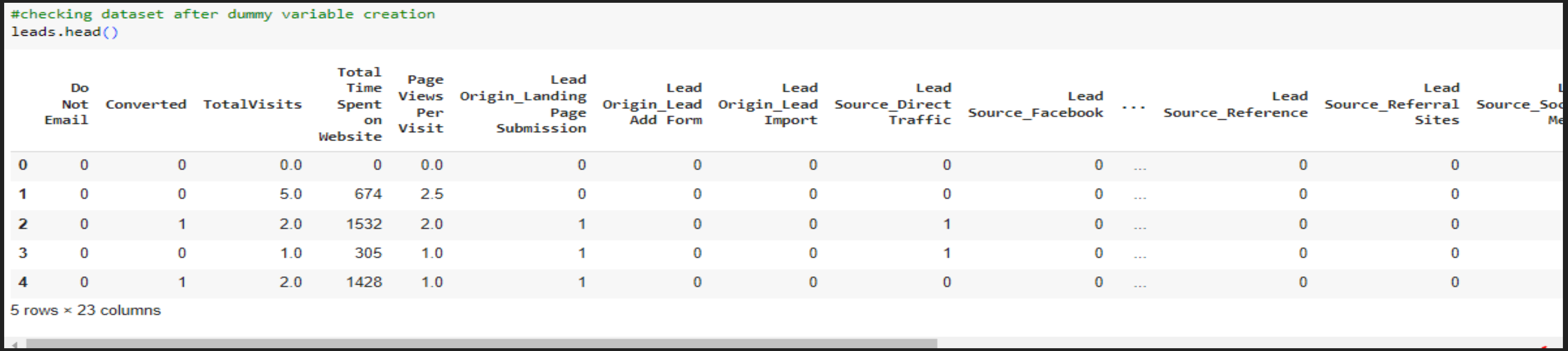
- Converted variable between 0 and 1



```
leads.head()
```

	Lead Origin	Lead Source	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	What is your current occupation
0	API	Olark Chat	0	0	0.0	0	0.0	Unemployed
1	API	Organic Search	0	0	5.0	674	2.5	Unemployed
2	Landing Page Submission	Direct Traffic	0	1	2.0	1532	2.0	Student
3	Landing Page Submission	Direct Traffic	0	0	1.0	305	1.0	Unemployed
4	Landing Page Submission	Google	0	1	2.0	1428	1.0	Unemployed

- Created dummy variable for categorical variable



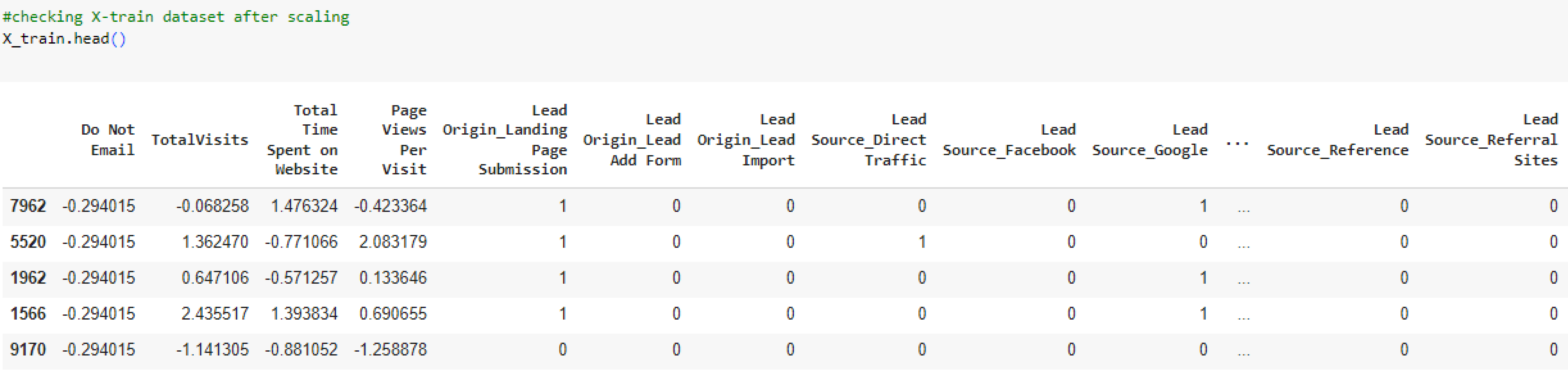
```
#checking dataset after dummy variable creation
leads.head()
```

	Do Not Email	Converted	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_Reference	Lead Source_Referral Sites	L Source_Soc Me
0	0	0	0.0	0	0.0	0	0	0	0	0	...	0	0	
1	0	0	5.0	674	2.5	0	0	0	0	0	...	0	0	
2	0	1	2.0	1532	2.0	1	0	0	1	0	...	0	0	
3	0	0	1.0	305	1.0	1	0	0	1	0	...	0	0	
4	0	1	2.0	1428	1.0	1	0	0	0	0	...	0	0	

5 rows × 23 columns

Feature Scaling and Splitting Train and Test

- Feature scaling of Numeric Data
- Splitting data into Train and Test

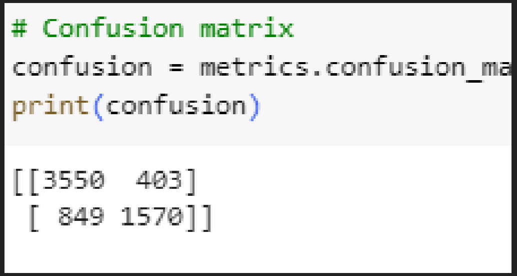


```
#checking X-train dataset after scaling
X_train.head()
```

	Do Not Email	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_Reference	Lead Source_Referral Sites
7962	-0.294015	-0.068258	1.476324	-0.423364	1	0	0	0	0	1	...	0	0
5520	-0.294015	1.362470	-0.771066	2.083179	1	0	0	1	0	0	...	0	0
1962	-0.294015	0.647106	-0.571257	0.133646	1	0	0	0	0	1	...	0	0
1566	-0.294015	2.435517	1.393834	0.690655	1	0	0	0	0	1	...	0	0
9170	-0.294015	-1.141305	-0.881052	-1.258878	0	0	0	0	0	0	...	0	0

Model Building

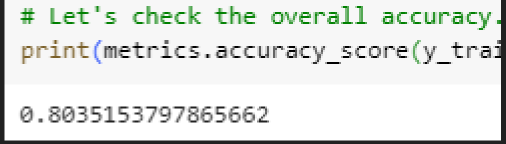
- Feature selection using RFE
- Determined Optimal model using Logistic regression
- Calculated accuracy ,sensitivity ,precision and recall and evaluate model



```
# Confusion matrix
confusion = metrics.confusion_ma
print(confusion)

[[3550  403]
 [ 849 1570]]
```

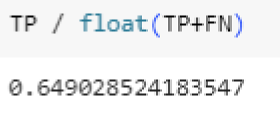
Confusion Matrix



```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_trai

0.8035153797865662
```

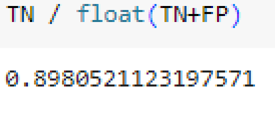
Accuracy



```
TP / float(TP+FN)

0.649028524183547
```

Precision



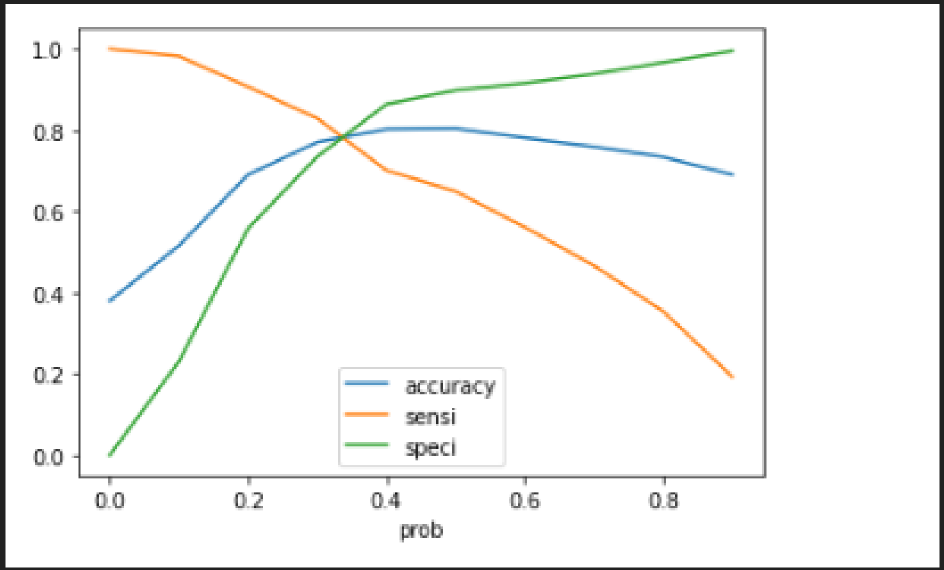
```
TN / float(TN+FP)

0.8980521123197571
```

Recall

Model Evaluation-Sensitivity and Specificity on Train Data Set

Graph depicts an optimal cut-off of 0.37 bases on Accuracy,Sensitivity,Specificity

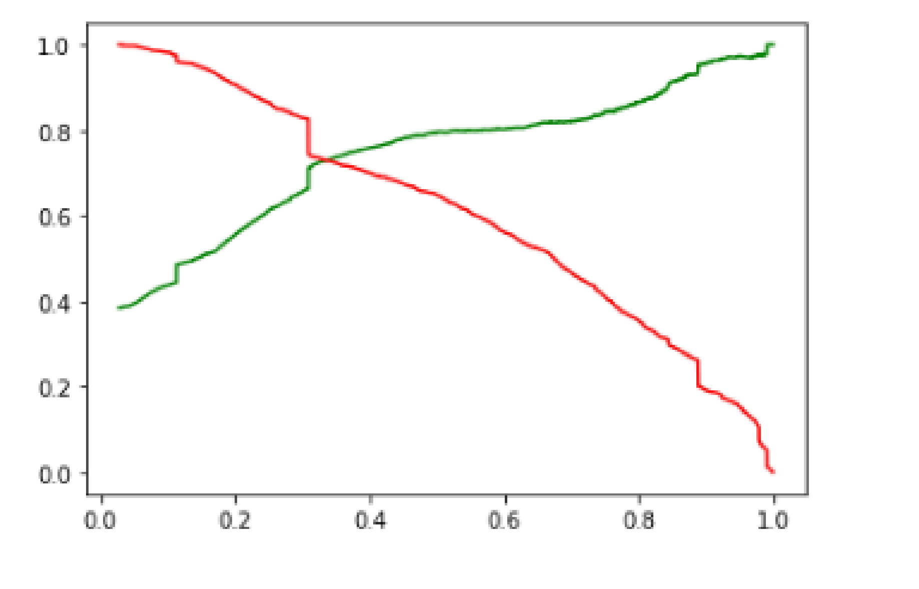


Accuracy : 78%
Sensitivity: 82%
Specificity: 76%

Model Evaluation

Precision and Recall on Train dataset

The graph depicts optimal cut-off of 0.42 based on precision and recall



Precision: 79%
Recall : 65%

Model Evaluation

Sensitivity and Specificity on Test Dataset

- ❑Accuracy: 78%
- ❑Sensitivity: 80.8%
- ❑Specificity: 76.5%

Result

- Accuracy, Sensitivity and Specificity values of training and test set are close to training set
- Accuracy, Sensitivity and Specificity values of training set are 79 %,82 %,76% Respectively

  Accuracy, sensitivity & Specificity values of test are 78%, 81%,76% Respectively
- Conversion rate for Train & Test Dataset is 82.7% & 80.8% Respectively
- We have done the prediction on the test set using cut off threshold from sensitivity & specificity metrics

Conclusion

- While we have checked both sensitivity-specificity as well as Precision & recall metrics, we have considered the optimal cut off based on sensitivity & specificity for calculating the final prediction
- Accuracy, Sensitivity & specificity values of test set are around 78%,81 %,76% which are approximately closer to Values calculated using Trained Data Set
- Lead Score Calculated for the conversion rate final model on Train & Test dataset is 82.7% 8.80.8% respectively.
- Hence, Overall Model seems to be Good

Summary

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (ie, educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion. First, sort out the best prospects from the leads you have generated. "Total Visits", "Total Time Spent on Website", "Page Views Per Visit' which contribute most towards the probability of a lead getting converted. Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies. Monitor each lead carefully so that you can tailor the information you send to them. Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects. Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them. Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses