The Lead Scoring Case Study on X_Education

Business Problem

- X Education sells online courses to industry professionals. Everyday many professionals who are interested in the courses land on their website and browse for courses. These users are classified as potential leads to take the course. Currently the conversion rate is very poor. The company wishes to identify the most potential leads, also known as ‘Hot Leads’. With the Hot Leads, the Sales team will focus more on communicating with the them rather than making calls to everyone.

Data Understanding

- Total number of rows -> 9240
- Total number of columns -> 37
- Few redundant features are identified at glance: “Prospect ID” and “Lead Number”
- Few columns have a value called ‘Select’ which has been assumed as due to not selecting any option by users, so imputed by Null values to get original count of Null values over the whole dataset.

- Numerical columns are : “TotalVisits”, “Total Time Spent on Website”, “Page Views Per Visit”
- Target feature is ‘Converted’ having two classes ‘No’(0) with 61.46% and ‘Yes’(1) with 38.53%. Small imbalance is noticed here.

Exploratory Data Analysis (Univariate and Bi-Variate)

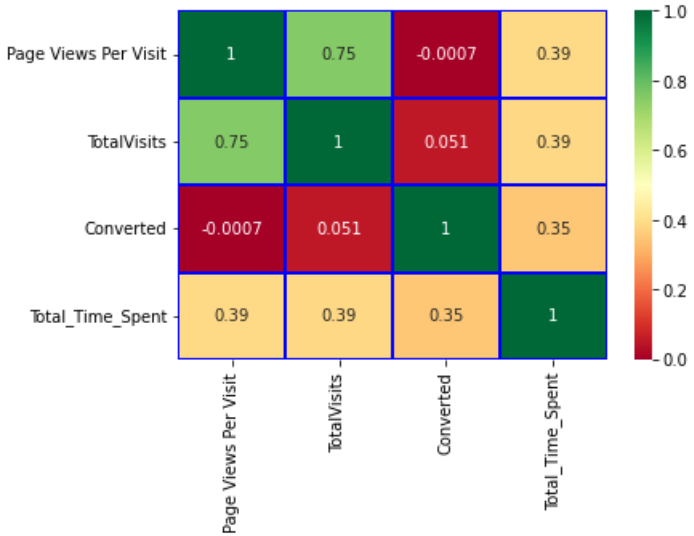
Numerical Variables	Categorical variables
• All the Numerical independent variables are skewed in nature and have outliers except “Total Time Spent on Website” and outlier percentage was below 5% of whole dataset. • All the rows having outliers were dropped from the dataset.	• Visualization made for each categories of each attribute to check the total count of target classes. • The inferences are given in next slide.

Categorical Feature Interpretation

- The Converted leads have less count for total visit then non-converted one.
- The average time spent visited on website by converted leads are higher than that of non-converted.
- The total number of pages viewed in every visit by the non-converted leads is higher than that of converted leads.
- The leads who have opened email has high percentage of non-conversion
- Leads with SMS Activity has high percentage of conversion among all.
- ‘Email Opened’ also has high percentage of conversion. Overall count of “Email Opened” and “SMS Sent” are pretty high.
- maximum leads have come from 'Olark Chat', 'Organic Search', 'Direct Traffic' and 'Google'. Very few from 'Referral Sites','Welingak Website', 'Reference' and 'Facebook'. Rest are very negligible.
- we can notice here 'Google' and 'google' - these same things are present in ‘Lead Source’ variable, but with different spellings. So we can convert these into same.
- Reference and Welingak Website - these two sources have higher conversion chance
- The categories which have very less count of leads irrespective of conversion , were omitted from the dataset.
- We can say that for many leads , the conversion didn't happen and they are just not interested to take the call , as a result the 'Ringing' category count is so high approximately 1155
- for 236 non- converted leads it has been found that their phone is switched off.
- 1960 leads have converted who have reverted after reading the mail.
- 496 leads are interested in other courses.
- 462 leads are already student - as a result the didn't turn up.

Checking Correlation

- In the dataset we checked the correlation among all the numerical columns. From there we found that total independent variables “TotalVisits” and “Page Views Per Visit” are multicollinear and that coefficient is 0.75.
- The target variable “Converted” has correlation coefficient 0.35 with “Total_Time_Spent” variable. The figure is given below:



Dummy Encoding (Preprocessing of Categorical Variables) and Train-Test Split

- Before dummy encoding the shape of the dataset was consisting 8517 rows and 12 columns after all feature elimination after EDA and insights analysis.
- After the dummy encoding the shape of the dataset came with 71 columns and 8517 rows.
- Post this the whole dataset was divided into train and test with 70% and 30% weightage accordingly.
- As the number of columns became 71 , so we ran recursive Feature Elimination(RFE) process to understand important features. The number of features were set at 20 in RFE process.
- The 20 features after RFE are shown in the screenshot in the next slide:

Selected features after RFE

- The 20 features are below:

	feature	feature_ranking	feature support
0	TotalVisits	1	True
1	Total_Time_Spent	1	True
2	Page Views Per Visit	1	True
7	Lead Source_Facebook	1	True
19	Lead Source_Welingak Website	1	True
21	Do Not Email_Yes	1	True
27	Last Activity_Olark Chat Conversation	1	True
29	Last Activity_SMS Sent	1	True
42	Current_Occupation_Other	1	True
43	Current_Occupation_Student	1	True
44	Current_Occupation_Unemployed	1	True
45	Current_Occupation_Working Professional	1	True
46	Tags_Busy	1	True
47	Tags_Closed by Horizzon	1	True
52	Tags_Interested in Next batch	1	True
55	Tags_Lost to EINS	1	True
57	Tags_Mobile Number Issue	1	True
58	Tags_Not Mentioned	1	True
62	Tags_Ringing	1	True
67	Tags_Will revert after reading the email	1	True

- Though Specialization related seemed to be one of the most important variables, those were eliminated by RFE process.
- We continued modelling with these 20 features.
- Current_Occupation is one of the most important variables and the dummy variables which were created from this seemed to be selected by RFE.

Modelling

- As it's classification problem , we started to build the statistical model 'Logistic Regression'.
- In total 4 models were built.
- After running the first model , it was found that only one variable is present with p- value with > 0.05, which is: Tags_Interested in Next batch .
- After rejecting tis variable we reran Logistic Regression Model, where all features seemed to be important. Then we tried to check Multicollinearity using Variance Inflation Factor(VIF) and we chose the threshold as 3. Whichever variable has VIF greater than 3, was eliminated from the analysis. Three variables were eliminated in this process which are: "Page Vies Per Visit"(6.28), "TotalVisists"(5.61), "Current_Occupation_Unemployed"(3.53).
- After this the total number of features became 16.
- In the third Model we found 4 variables are present with p-value > 0.05 which are: "Lead Source_Facebook", "Current_Occupation_Other", "Current_Occupation_Student", "Current_Occupation_Working_Professional".
- By intuition all these seem to be very important but "Current_Occupation_Working_Professional" seemed to be most important one because it actually relates to the business problem as the X_Education organization want to sell its courses to the industry professionals. So Except this variable all three were dropped from the dataset and the features of the dataset came down to 13.
- We again ran Logistic Regression Model and that model was chosen for the further analysis.

Prediction on Train Set

- The result of prediction on the Train set is given below:

	Actual	Predicted_Prob
0	1	0.995541
1	0	0.535533
2	0	0.016526
3	1	0.992071
4	0	0.437571

This prediction was done by default on the basis of threshold of 0.5, So we rounded this predicted numbers to 1 and 0 based on that.

	Actual	Predicted_Prob	predicted
0	1	0.995541	1
1	0	0.535533	1
2	0	0.016526	0
3	1	0.992071	1
4	0	0.437571	0

The Confusion Matrix is given below:

```
[[3539  164]
 [ 358 1900]]
False Positive rate :  4.428841479881177
correctly specified percentage :  91.24308002013085
Incorrectly specified percentage :  8.75691997986915
Positive Prediction rate(Precision) :  92.05426356589147
Negative Prediction Rate :  90.81344624069797
Sensitivity(Recall) : 84.1452612931798
Specificity : 95.57115852011883
```

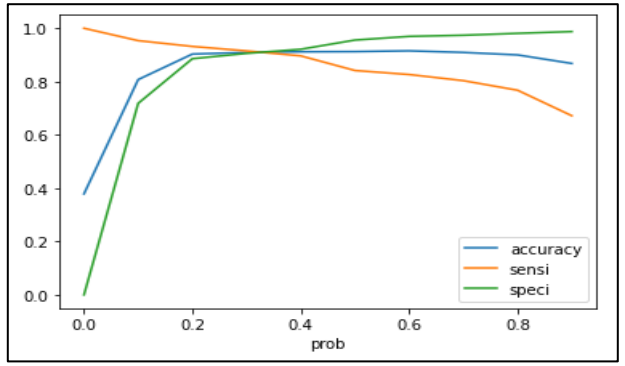
- **Accuracy is 91.24%** -> Which implies that approximately 91.24% of the predictions made by the model were correct.
- **Incorrectly specified percentage is 8.75%** -> Which implies that in 8.75% cases model's prediction is not aligning with actual labels
- **Precision or Positive Prediction Rate is 92.05%** -> Wich implies that, when a model predicted a positive outcome, it was correct almost 92% cases. Also Precision gives the reliability of positive prediction.
- **Negative Prediction Rate is 90.81%** -> When the model predicted negative outcome, it was correct almost in 91% cases.
- **Sensitivity or Recall is 84.14%** -> Which implies that approximately 84.14% of all actual positives were classified correctly. This metric is important when we want to focus only on identifying actual positive classes.
- **Specificity is 95.57%** -> Which implies that approximately 95.57% of all actual negatives were classified correctly. A HIgh Specificty indicates that model is effective in identifying negative classes correctly.

- The whole interpretation shows that the model is pretty good and it's quite capable in classifying the leads into two classes (1 and 0).
- Post that we tried to identify the optimum threshold value. Basically , for several values between 0 to 1 we tried to check the round off the predicted probability values. Thereafter for those values we derived accuracy, sensitivity and specificity. These graphs were plotted. The intersection point was taken as our optimum point for further analysis.

Finding the optimum Threshold

- For finding the threshold we took some numbers as threshold from 0 and 1 and tried to analyse Accuracy , Sensitivity and Specificity. The intersection point was chosen as the cut-off for further analysis.

	prob	accuracy	sensi	speci
0.0	0.0	0.378796	1.000000	0.000000
0.1	0.1	0.807415	0.953499	0.718336
0.2	0.2	0.903204	0.931798	0.885768
0.3	0.3	0.909243	0.914969	0.905752
0.4	0.4	0.911760	0.896368	0.921145
0.5	0.5	0.912431	0.841453	0.955712
0.6	0.6	0.915283	0.826395	0.969484
0.7	0.7	0.909243	0.803366	0.973805
0.8	0.8	0.900017	0.767493	0.980826
0.9	0.9	0.867975	0.671833	0.987578



- Confusion Matrix for the same is below:

```
[[3354  349]
 [ 192 2066]]
False Positive rate :  9.424790710234944
correctly specified percentage :  90.92434155343064
Incorrectly specified percentage :  9.075658446569367
Positive Prediction rate(Precision) :  85.54865424430642
Negative Prediction Rate :  94.585448392555
Sensitivity(Recall) : 91.49689991142604
Specificity : 90.57520928976506
```

Which can be interpreted as the model can provide accuracy of approximately 91% and also Sensitivity is found to be 91.5% and specificity is 91%.

The same model performed very well on the test set i.e, any unknown data. The performance metrics and their interpretations are explained in the next slide.

Performance on Test Set

- The predictions on Test Set are below:

	Actual	Predicted	final_predicted
0	0	0.357001	1
1	0	0.099690	0
2	0	0.805370	1
3	1	0.242719	0
4	0	0.057220	0
...	...	...	...
2551	0	0.176259	0
2552	0	0.037948	0
2553	1	0.417988	1
2554	0	0.012295	0
2555	0	0.083114	0

2556 rows × 3 columns

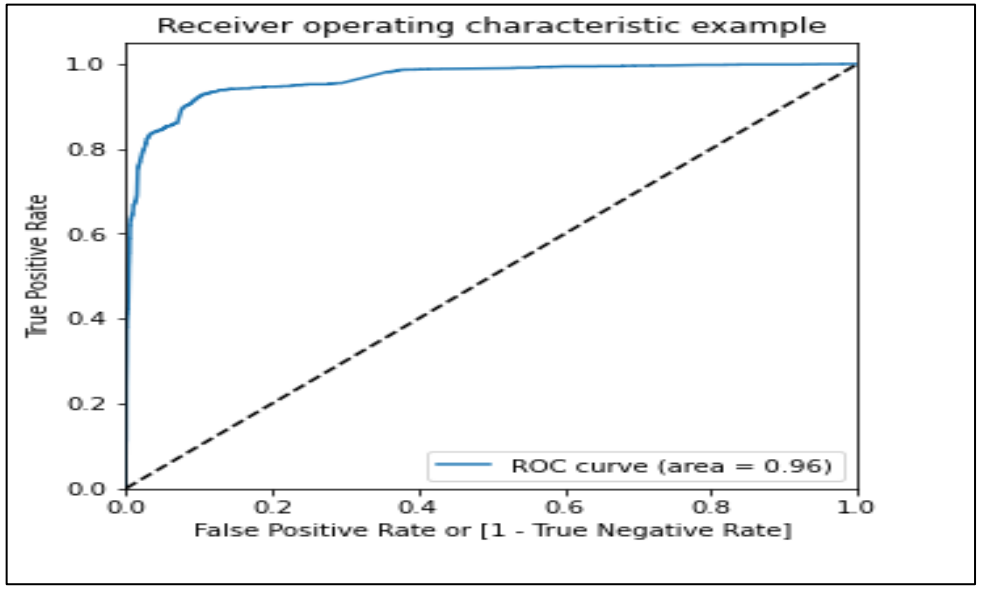
- The Confusion Matrix and its interpretation is attached below:

```
[[1429  177]
 [  75  875]]
False Positive rate :  11.021170610211705
correctly specified percentage :  90.14084507042254
Incorrectly specified percentage :  9.859154929577464
Positive Prediction rate(Precision) :  83.17490494296578
Negative Prediction Rate :  95.01329787234043
Sensitivity(Recall) : 92.10526315789474
Specificity : 88.97882938978829
```

- For the test data or unknown data the model can provide 90.14% accuracy i.e, approximately 90.14% predictions made by the model are correct.
- 9.85% predictions by the model are erroneous.
- Precision is 83.17% i.e, when the model predicts something positive, it's found to be correct almost for 83.17%.
- Negative Prediction rate is 95.1% which says that out of all predicted negative by the model, approximately 95.1% is correct.
- Sensitivity or Recall is 92.10% which implies that the model can correctly identify approximately 92.10% Actual Positives out of all actual positives.
- Specificity is 88.97% which implies that the model can correctly identify approximately 88.97% Actual Negatives out of all actual negatives.

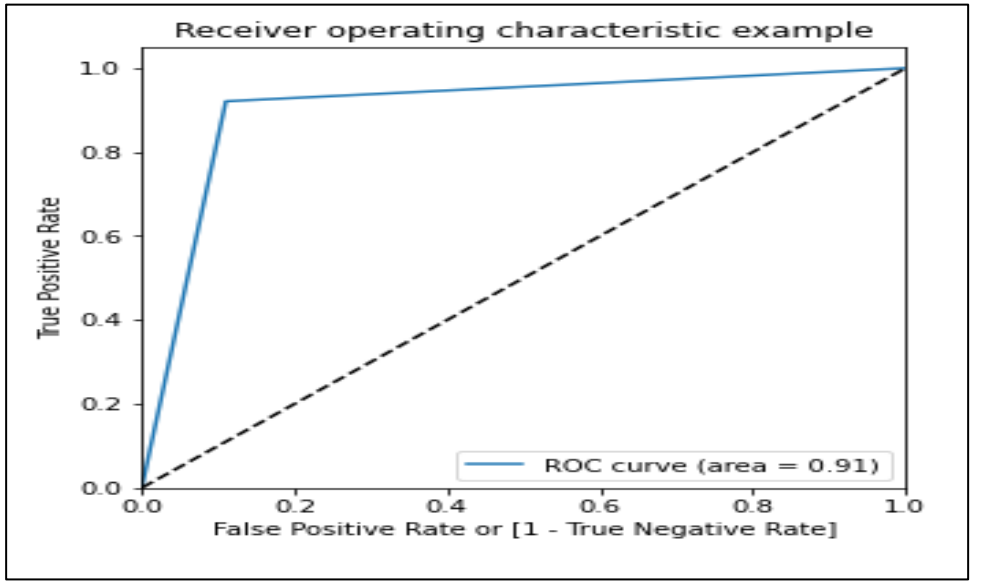
AUC-ROC Curve for Train Set and Test Set

- Train set AUC-ROC Curve is below:



The Train set ROC score is 0.96

- Test set AUC-ROC Curve is below:



The Train set ROC score is 0.91

Conclusion:

- So, we can say that we have built a model with 91.24% Accuracy which can provide approx 90% Accuracy on test set.
- Sensitivity or Recall or True Positive Prediction Rate has been found as 84%-92% for test set and for train set also it gave the same range which implies that approximately 84% to 92% cases it can classify the actual positive classes correctly.
- Specificity is also pretty good(~ 95%), which indicates approximately in 95% cases it can classify negative classes correctly.
- The Statsmodel and Scikit-learn model shows small amount of variation in their results.
- As per the business problem, it's very imporatnt to identify the potential leads, So here, identifying True Positive or metric 'Recall' is most important.
- So, This Logistic Regression model is a good model for this type of business problem.

Thank You