

LEAD SCORING – CASE STUDY



Submitted by:

Anmol Mehta

K Chandrahas

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.



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, 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 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. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

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 - 2

- **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 **MinMax 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

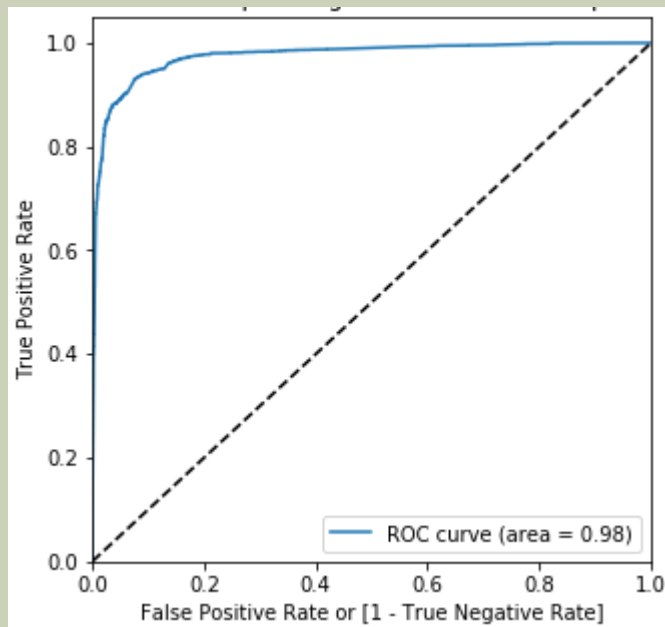
- After training the model on Train set, predictions were made and their evaluation metrics were recorded.
- We kept a cut-off of 80% while making the predictions.
- Confusion Matrix was created :

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

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

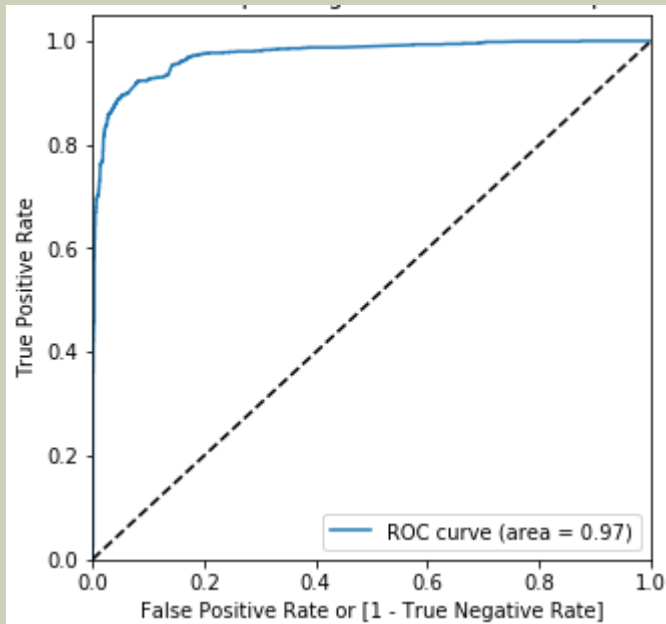
- After training the model on Train set, the predictions were made on the Test data set and their evaluation metrics were recorded.
- We kept a cut-off of 80% while making the predictions.
- Confusion Matrix was created :

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.