Lead Scoring

CASE STUDY

Table of Contents

Problem Statement

Education company X Education sells online courses to industry professionals.

Lead Generation and Conversion

Markets on several websites, search engines like Google, past referrals.
Once leads are acquired, sales team start making calls, writing emails, etc.
Some leads get converted while most do not.
Typical lead conversion rate is around 30%.
Lead conversion rate is very poor.
To make this more efficient, company wishes to identify the most potential leads; ‘Hot Leads’.

Significance of solving the problem

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.

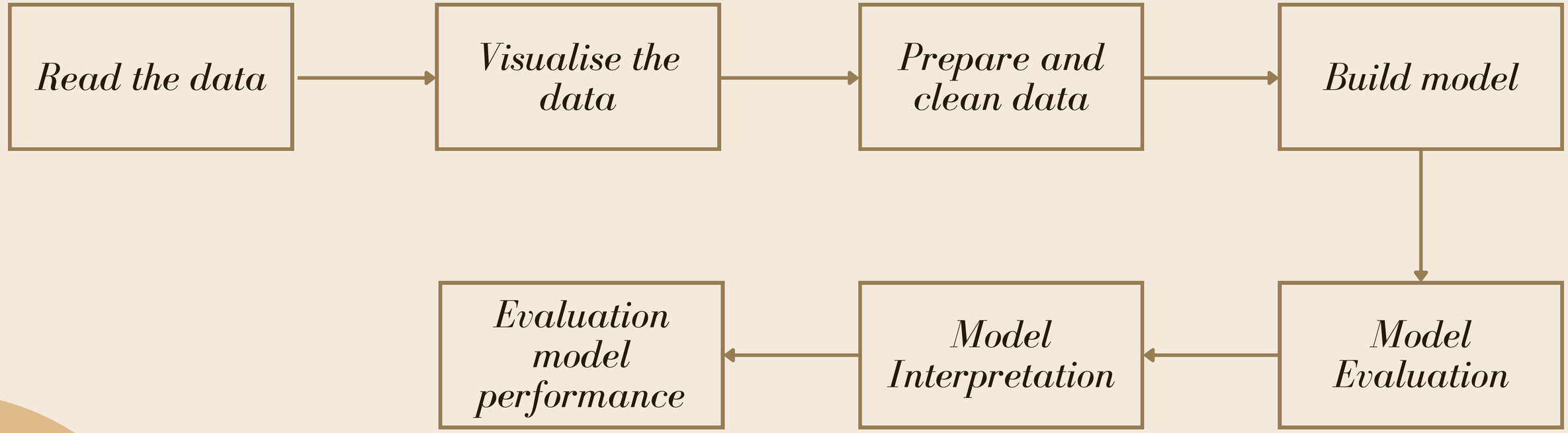
Business Context

Objective:
Identify leads that are most likely to convert into paying customers

Expected outcomes:
Build a logistic regression model to assign a lead score between 0 and 100.
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.

Impact on Business:
Increase conversion and change startegy based on how many persons are available in the sales team.

Analysis Approach



Data Understanding

Data Description

- The target variable is ‘Converted’. 0 means lead did not convert, while 1 means converted
- Of the remaining 36 variables, 2 are unique ID variables, 8 are numeric and rest are categorical variables.
- Categorical variables have 2 - 20 categories.
- Some of the categorical variables have a category called ‘Select’ which is a default value needs to be treated.
- Variables ‘Asymmetrique Activity Index’, ‘Asymmetrique Profile Index’, ‘Asymmetrique Activity Score’, ‘Asymmetrique Profile Score’ had similar missing values and since all relayed similar information, we retained only one of the variables.

Exploratory Data Analysis

EDA

- Conducted univariate analysis to identify trends in the data.
- Conducted bivariate analysis to identify how various categories are related to the target variable; ‘Converted’.

Data Preparation

Data Cleaning

Handling Missing Values:

- For variables which had null values lower than 5%, rows containg null values were deleted.
- Variables that had ‘Select’ as a value were converted to null.
- For variables which had null values greater than 5% null values were converted to ‘Unknown’.

Outlier treatment:

- Outliers were not particularly treated, instead the min-max scaler was used to standardise all variable range. This took care of the outliers.

Data Preparation

Feature Engineering

Creating new features:

- For categorical variables, label encoding method was adopted. This is because there were numerous categories and a substantial number of categorical variables. This would increase the number of variables significantly loading the model.

Feature transformation:

- The min-max scaler was used to standardise all numeric variables.

Splitting data:

- Data was split into train and test datasets; 70% for training the model.

Modelling Approach

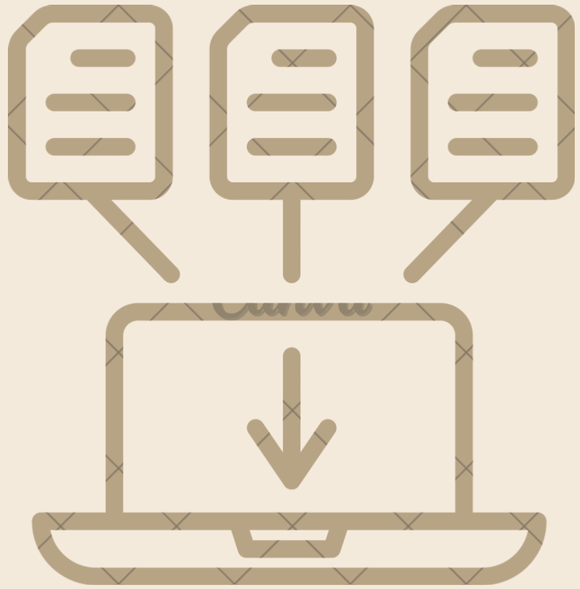
Objective:
Identify leads that are most likely to convert into paying customers

Choice of Model:
Logistic regression will be a suitable model for this business case. Logistic regression is a supervised machine learning model used for classification, i.e, to predict the probability that an instance belongs to a given class or not. it helps to make predictions for categorical variables.

Validate the model:
To validate the model, confusion matrix and ROC curve has been used.

Model Building

Training Process



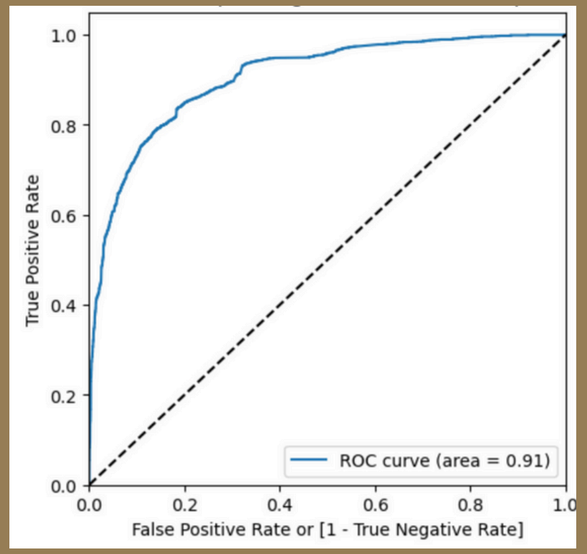
Feature Selection:

- Logistic Regression model was run using sklearn
- With the help of RFE (Recursive Feature Elimination) top 15 variables were shortlisted.

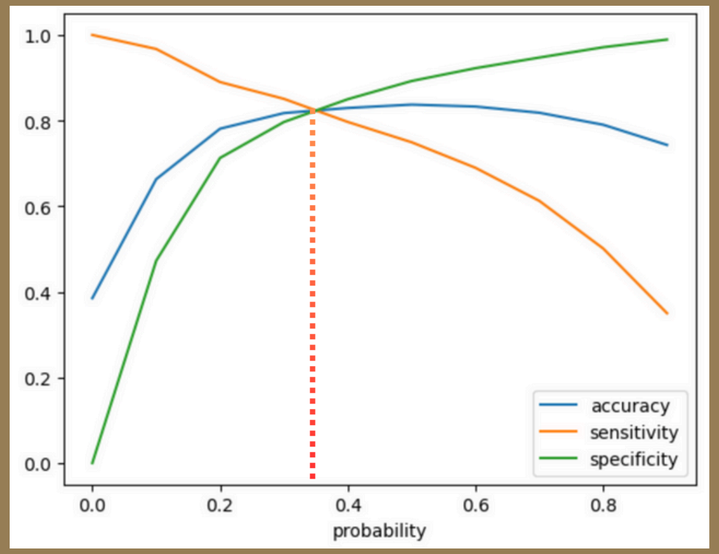
Hyperparameter Tuning:

- On successive models, VIF (Variance Inflation Factor) and p value were used to eliminate variables that either had high multicollinearity (VIF >5) or were statistically insignificant (p value > 0.05).

Model Evaluation



ROC curve of the model



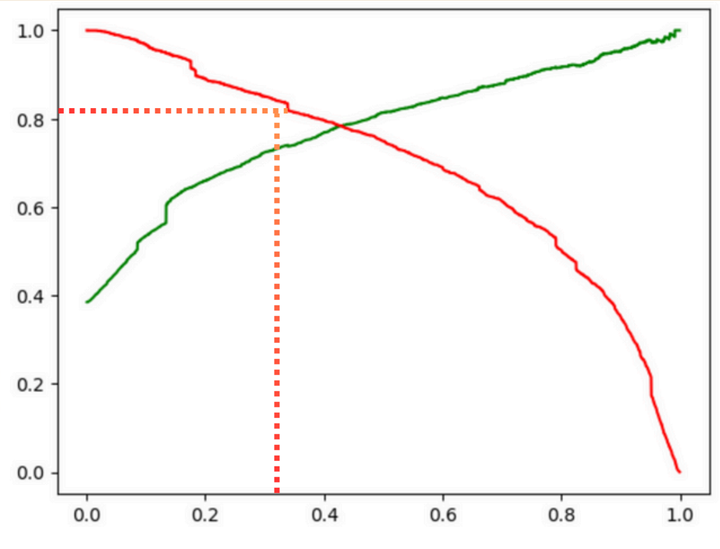
Sensitivity, Specificity, Accuracy plot of the model

From the Sensitivity, Specificity and Accuracy plot (for cut-off threshold 0.5) on the left, the optimum threshold of 0.35 was obtained.

Performance Metrics with cut-off as 0.35

- Accuracy = 81.8%
- Sensitivity = 81.4%
- Specificity = 82.0%

Model Evaluation



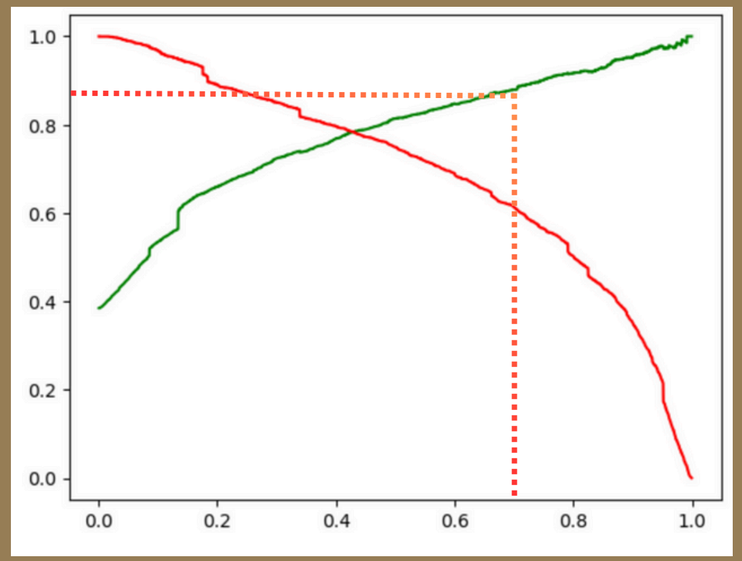
Precision Recall trade-off curve
Green: Precision
Red: Recall

For the business case where the company wants to make calls to as many potential leads, it is important to detect as many potential 'hot leads'.
To do this we need to increase recall and thus reduce the cut-off threshold. Since the company target is to reach 80% conversion, we will take the cut-off of 0.3

Performance Metrics with cut-off as 0.3

- Accuracy = 81.7%
- Sensitivity/ Recall = 85.0%
- Precision = 72.0%

Model Evaluation



Precision Recall trade-off curve
Green: Precision
Red: Recall

For the business case where the company reaches its target before the end of the quarter and wants to deploy its resources elsewhere and make less phone calls, we need to ensure that the leaads that the model is detecting as ‘hot leads’ have a high potential of converting.
To do this we need to increase precision and thus increase the cut-off threshold.
We can make the cut-off between 0.6-0.8.

Model Interpretation

Top Features

1. Time spent on website (positively affects)
2. Pages viewed per visit (negatively affects)
3. Do not email
4. Lead Quality
5. Lead Origin

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1036	0.197	-5.592	0.000	-1.490	-0.717
Lead Origin	0.5277	0.075	7.007	0.000	0.380	0.675
Lead Source	0.1451	0.014	10.032	0.000	0.117	0.173
Do Not Email	-1.4634	0.177	-8.263	0.000	-1.810	-1.116
Total Time Spent on Website	4.2489	0.169	25.100	0.000	3.917	4.581
Page Views Per Visit	-4.7586	0.505	-9.414	0.000	-5.749	-3.768
Last Activity	0.1253	0.010	12.001	0.000	0.105	0.146
What matters most to you in choosing a course	-0.2742	0.032	-8.447	0.000	-0.338	-0.211
Tags	0.0759	0.006	12.047	0.000	0.064	0.088
Lead Quality	-0.9228	0.036	-25.393	0.000	-0.994	-0.852
City	0.0643	0.020	3.152	0.002	0.024	0.104
A free copy of Mastering The Interview	-0.2953	0.094	-3.139	0.002	-0.480	-0.111

Model Performance

Model Metrics on Test Dataset

Metrics	Train Set	Test Set
Accuracy	81.7%	79.7%
Precision	72.0%	67.0%
Recall	85.0%	86.5%

Accuracy has marginally dropped

Since Precision is not a priority in the BAU case, it's okay if it drops

Recall is better on this test set

Conclusion

Addressing the problem of lead generation and conversion for X Education is crucial for ***enhancing the efficiency of the sales process*** and ***improving overall business performance***. The current lead conversion rate of approximately 30% indicates a significant opportunity for improvement. By identifying and targeting 'Hot Leads,' the company can optimize its resources, allowing the sales team to focus their efforts on the most promising prospects.

Business Impact:

The significance of solving this problem cannot be overstated. Successfully pinpointing potential leads will not only increase the lead conversion rate but also ***reduce the time and effort*** wasted on less likely conversions. Furthermore, a higher lead conversion rate will ***reinforce X Education's market position***, enabling the company to better serve industry professionals with their online courses. Ultimately, this strategy will contribute to ***sustained growth and success*** for X Education, ensuring that they ***remain competitive*** in the rapidly evolving education market.

Thank You