Lead scoring case study

By Sayyad Munwwarali
20 June 2023

Problem Statement

X Education sells online courses to professionals and receives leads through website visits, form submissions, and referrals. However, their lead conversion rate is low, and they want to improve it by identifying the most potential leads, or 'Hot Leads.' They aim to build a model that assigns lead scores to increase the chances of converting leads into paying customers, with a target conversion rate of 80%.

Path to solution

1) Data Loading
2) Data Exploration a.k.a Exploratory Data Analysis
3) Preprocessing
4) Feature Engineering
5) Outlier Analysis
6) Model Building
7) Model Performance Benchmarking
8) Model Performance Evaluation
9) Cross Validation
10) Model Diagnosis Using ROC AUC Curve, Precision-Recall Curve

Data loading and EDA

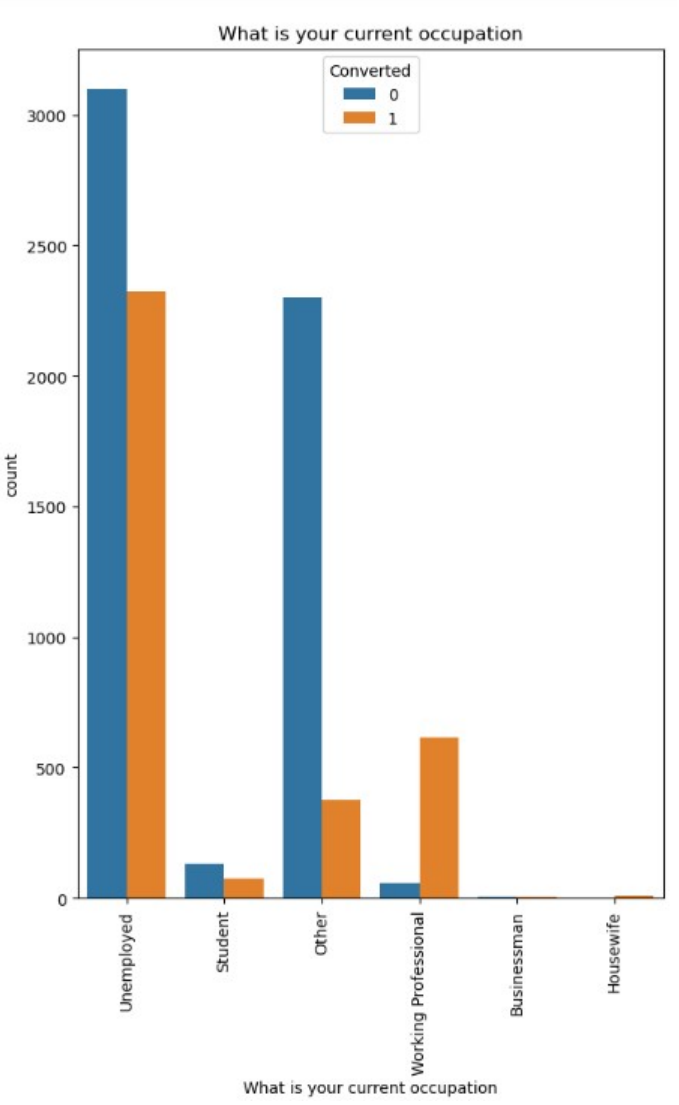
WE are using pandas library along with matplotlib and seaborn to read and perform preliminary analysis on the dataset.

During EDA observed that there is higher conversion rate on for leads who have spent more time on the website and along with this if their lead origin is API or Landing page submission.
People who approach upfront or reach out via email are also very strong leads

EDA Observation

Among other observations this one stands out.

While the working professionals have heigher conversion ratio unemployed leads have highest number of conversions.



Final Model

We are using logistic regression model to classify the leads as hot or not by their lead score varying up to 100

Since the final goal is to classify whether the lead is hot or not, this becomes a classification problem.
For this problem logistic regression is a good starting point.
Out of all the features final model has top influential features with permissible VIF and p value.
With this model our threshold for classification is at 0.26.

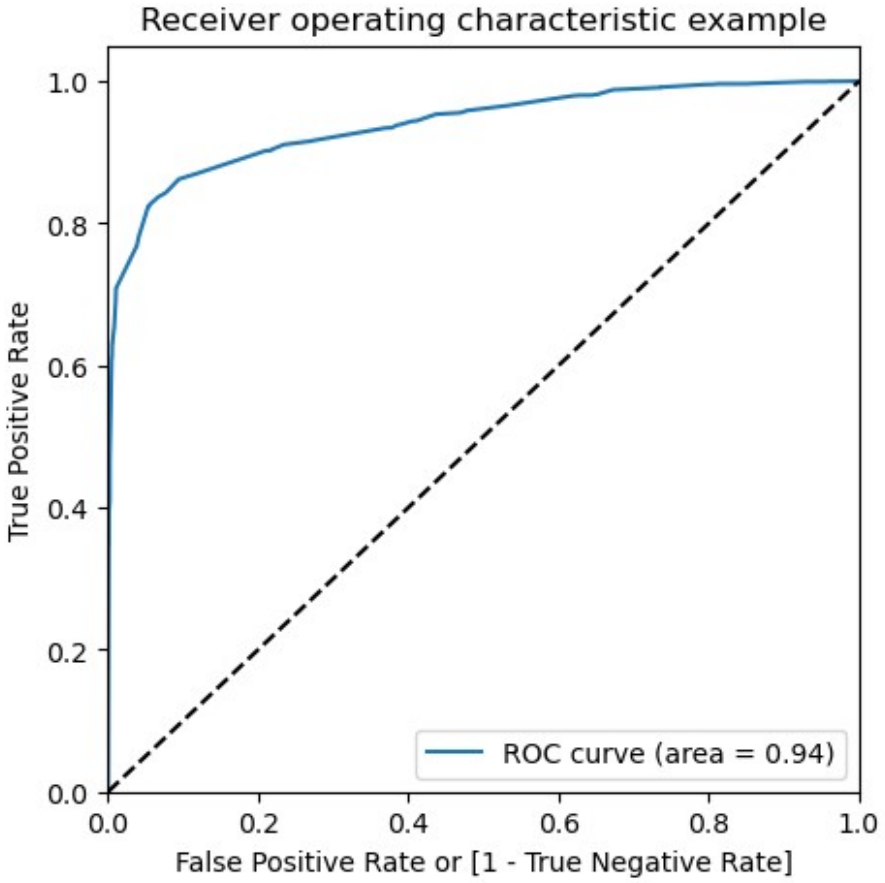
Train data metrics: Accuracy : 88.89% Sensitivity : 86.23% Specificity : 90.49%

Test data metrics: Accuracy : 83.06% Sensitivity : 90.17% Specificity : 78.61%

ROC Curve

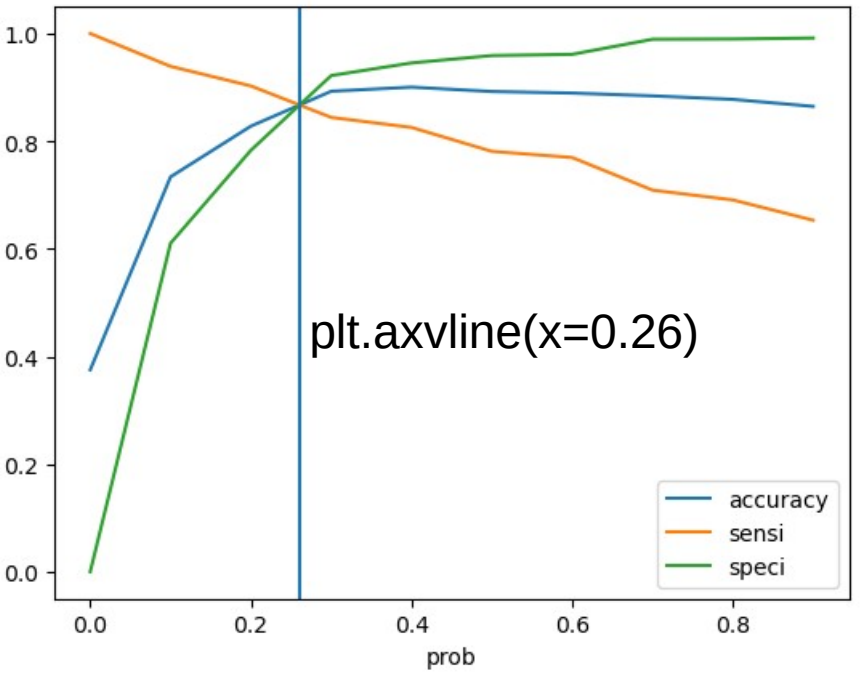
Area under curve is 0.94

That means the model is highly capable of distinguishing between classes.



Optimum cut-off

0.26 is the optimum point to take it as a cut-off probability



Thank you