Lead Generation for X Education

By Linus Jen

Table of Contents

Problem Statement

01

Improving conversion rates to increase sales

Problem

The CEO of X Education, a company selling online courses to industry professionals, wants to improve sales by improving lead quality. He has tasked the team to improve their conversion rate from 35% to 80%.

Solution

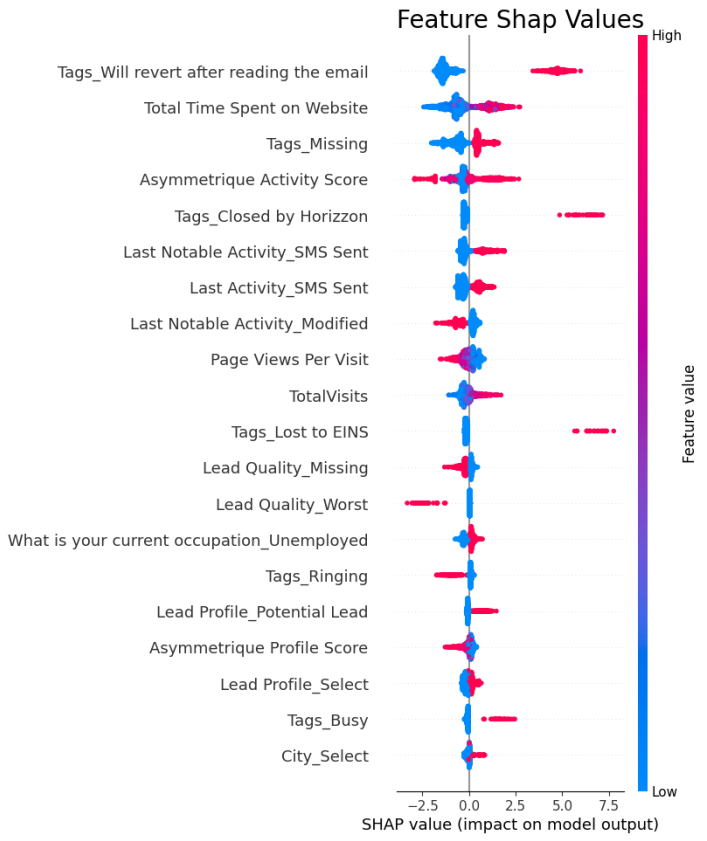
02

Apply ML model to target prospects

Solution

Given the survey results for those who have created an account with X Education and their decision to convert, I built a classification model with high metrics (95% accuracy, recall, and F1).

Feature Importance



Shap values give insight to black box models by showing prediction tendencies

- Tags describe attributes about a potential customer
- Last (Notable) Activity shows the last interaction a customer had with the website or company
- Customer actions (time spent on website, activity scores, page views per visit, etc.) seem to dictate desire of product

Prospect Profile

- High interactions with website (spend lots of time on website, have high magnitude for asymmetrique activity scores, and visit our site often)
- Look for particular tags (those who would revert after reading the email, missing tags, closed by Horizzon(?), lost to EINS(?)
- Last (Notable) Activity of SMS sent by the user also tends to improve the likelihood of converting
- Avoid: those who have many page views per visit, missing / worst rated leads, or their Last Notable Activity was modified

Potential Growth

- Expand to other countries
  - People from India made up over 90% of observations
- Certain types of students (Dual Specialization and Lateral) have seen high number of conversions in small samples
  - Potentially do outreach to specialized schools? Sponsor hackathons, provide courses, etc.?

Technical Detail

03

Procedural review for non-stakeholders

Data Preprocessing

Removed 13 features with low variance

- EX: `Do Not Call` was filled with mostly False (2 True's)

Combined categories with low # of observations

- EX: `Countries` with under 50 total observations were aggregated into one group

Imputed `Missing` for categorical variables and 0 for numerical variables

- Missing data had unique proportions for conversions, as compared to other groups

SMOTE (oversampling minority classes) and undersampling

- Used to address class imbalances

Model Building

Final preprocessed data (after one-hot encoding): 108 columns

- Split data into train / validation / test sets of 70/15/15

XGBoost was the only model tested due to time constraints

- Accuracy, weighted precision all 0.95, weighted recall 0.93

Final model: XGBoost with feature selection (81 columns)

- Accuracy = 0.95, weighted precision 0.94, weighted recall 0.93

Application & Limitations

04

Pilot plan & assumptions

Pilot Plan

01

Treatments

Split all prospects into two groups – one following existing method, and another recommended by model

02

Compare

After trial period, run a two sample t-test to test of improvement of conversion rate

03

Decide

Determine whether or not to implement model

Limitations - Data

Dataset is limited to questionnaire responses *after* signing up to website. Assumes truthful responses

No data is present for characteristics of a prospect's online presence (interests, age, etc.)

Model only targets those who have already signed up to X Education – existing lead generation strategy not shared

Limitations - Procedure

Only XGBoost was tested in the modeling process (without hyperparameter tuning)

Given model success, did not change the threshold at which a prospect would be predicted as likely to convert

Difficult to test model generalizability when data is limited to people who have already signed up to the website