LE TRUNG KIEN - Lead Scoring Case Study - Data Science Program (Global) C12

SUMMARY REPORT

Problem statement:
X Education sells online courses. They recruit students from various channels including websites, search engines. Via marketing channels, they get a number of people who are interested in the provided courses, known as 'leads'. Their sales team need to convert these leads into actual students by different means (phone calls, sms, emails,…). The current conversion rate is around 30%. X Education wants to improve the rate to about 80%.

Solution: as a (or a team of) data scientist, we will build a model to help the company identify strategies to improve the lead conversion rate. We followed the steps below:

- **Step 1: Data cleaning & manipulation**

  Dropped unwanted columns:
  - We noticed a number of columns contain a large amount of null values. We also noticed a number of columns contain the 'Select' value, which is as good as null.
  - ➔ We converted 'Select' into nulls and check the columns. Then, we then dropped columns that have more than 45% of missing values (nulls).
  - We then removed columns where there is just 1 unique data value and columns containing all unique values as they do not contribute to in-depth analysis
  - After that, we checked the remaining columns and drop unwanted columns such as Tags (a lot of missing values and vague), City (skewed with a lot of missing values)…

  Dealt with missing values:
  - Depending on the nature of the columns, we filled missing values by imputing them with the most occurred values for categorical columns or replacing them with the mode for numerical columns.

  Other works:
  - Then we treated outliers, grouped low frequency values, mapped binary categorical values, and standardized values.

  After this stage, we had 10 data columns left, ready to be analyzed.

- **Step 2: Exploratory data analysis**
  - We checked data imbalance and found that only 38.54% of leads were successfully converted.
  - We then performed univariate analysis and found that:
    - The majority of leads come from "Landing Page Submission" (52.88%) and "API" (38.74%).
    - Nearly 90% of leads are unemployed.
    - Of the known specialization, people who worked in Finance Management, HRM, and Marketing Management have a higher tendency to be attracted by the courses provided by X Education.
    - Leads seem to be more keen on opening emails and sending sms.
  - We continued with performing bivariate analysis. The results were:
    - Landing Page Submission accounts for about 53% of the leads, while that of API and Lead Add Form are 29% and 8% respectively. The Conversion rates

of Landing Page Submission, API, and Lead Add Form are, therefore, 31%, 36%, and 88% accordingly.
    - Of all the sources, leads that come from Reference and Welingak Website are mostly sucessfully converted. Yet these two only account for 5% and 2% of all lead count. Google tops the contribution, generating 32% of leads. Direct Traffic, Olark Chat, and Organic Search follow with 28%, 19%, and 13% of lead count. Google is the better source among these 4 as it has a conversion rate of 41%.
    - Approximately 63% of potential customers who sent an SMS become an X student, the highest conversion rate based on Last Activity. Top lead contributors are groups who opened email (38%), sent SMS (30%), or participated in olark chat (11%). Olark Chat seems to be ineffective since only 9% of clients were converted.
    - Of the known categories, Finance Management (11%), HRM (9%), and Marketing Management (9.9%) are more significant compared to other groups. These groups' conversion rate are fairly high (>44%) with Marketing Management almost reached 50% of conversion.
    - Although the majority of leads come from people who do not have a job, only 34% of them participated in X's courses. Working Professionals account for just 8% of leads but this group's conversion rate is remarkably high (88%).
    - Distributing a free copy of Mastering The Interview does not seem to improve conversion rate as only 35% of the people who received Mastering the Interview are converted comparing with 40% for those who did not.

- **Step 3: Preparing data for model building**
    - We created dummy variables for categorical variables.
    - We then split the Train – Test sets at 70:30 ratio.
    - After that, we scaled numerical data with the MinMaxScaler

- **Step 4: Model building**
    - On the first model, we used RFE to select 15 most influential features.
    - On the second model, we removed variables with p-value > 0.05.
    - On the third model, we removed variables with VIF > 5.
    - We have our final model with 13 variables left.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6454
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2721.7
Date:	Tue, 31 Oct 2023	Deviance:	5443.5
Time:	21:31:39	Pearson chi2:	9.82e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3860
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5505	0.148	-10.473	0.000	-1.841	-1.260
TotalVisits	0.6484	0.174	3.730	0.000	0.308	0.989
Specialization_Others	-1.0735	0.122	-8.832	0.000	-1.312	-0.835
Last Activity_Olark Chat Conversation	-1.3647	0.169	-8.085	0.000	-1.695	-1.034
Last Activity_Had a Phone Conversation	2.0347	0.659	3.087	0.002	0.743	3.326
Last Activity_Email Bounced	-1.7902	0.295	-6.071	0.000	-2.368	-1.212
Last Activity_Converted to Lead	-1.0267	0.222	-4.635	0.000	-1.461	-0.593
Lead Source_Welingak Website	2.4715	0.749	3.298	0.001	1.003	3.940
Lead Source_Olark Chat	1.1915	0.131	9.111	0.000	0.935	1.448
Last Activity_SMS Sent	1.2342	0.074	16.720	0.000	1.090	1.379
Current Occupation_Working Professional	2.7089	0.191	14.205	0.000	2.335	3.083
Web Time	4.4132	0.161	27.336	0.000	4.097	4.730
Lead Origin_Lead Add Form	3.2095	0.208	15.418	0.000	2.802	3.618
Lead Origin_Landing Page Submission	-1.1317	0.126	-8.955	0.000	-1.379	-0.884

- **Step 5: Model evaluation**
  - We found the optimal cutoff point at 0.35.
  - Confusion matrix was made at the optimal cutoff point and retrieved an accuracy of approximately 81%. The sensitivity & specificity rates were 81%. Precision value was at 73% while Recall was 81%.
  - We performed precision-recall tradeoffs but sensitivity and specificity were dropped, so we sticked with the sensitivity-specificity view (cutoff point 0.35).

- **Step 6: Making prediction on test data**
  - We trained the data with our final model using 0.35 as cutoff point.
  - We had around 81% of accuracy, 75% sensitivity, and 85% specificity. Precision and recall were at 76% and 75% respectively.
  - This comes close to X Education's requirement of building a model aiming for the target lead conversion rate of around 80%.

- **Step 7: Giving conclusion and recommendations**

  The top 3 features that contribute to better prediciting hot leads are:
  - Web Time
  - Lead Origine_Lead Add Form
  - Current Occupation_Working Professional

  Our recommendations are:
  - X Education may want to invest in upgrading their website with new features, more interesting information as those who are attracted to the website have a higher tendency to become students.
  - The company may also want to have a marketing strategy focusing on working professionals.
  - Increasing budgets for Welingak Website, Olark Chat may help boosting hot leads.
  - Spend more on telephone-marketing and sms campaigns.
  - The landing page submission as well as email bounced should be improved.