Lead Scoring Case Study Analysis

Submitted By:
Samriddhi Vijayvargiya
Samiksha Kumbhare
Apphia Samuel

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.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, 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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
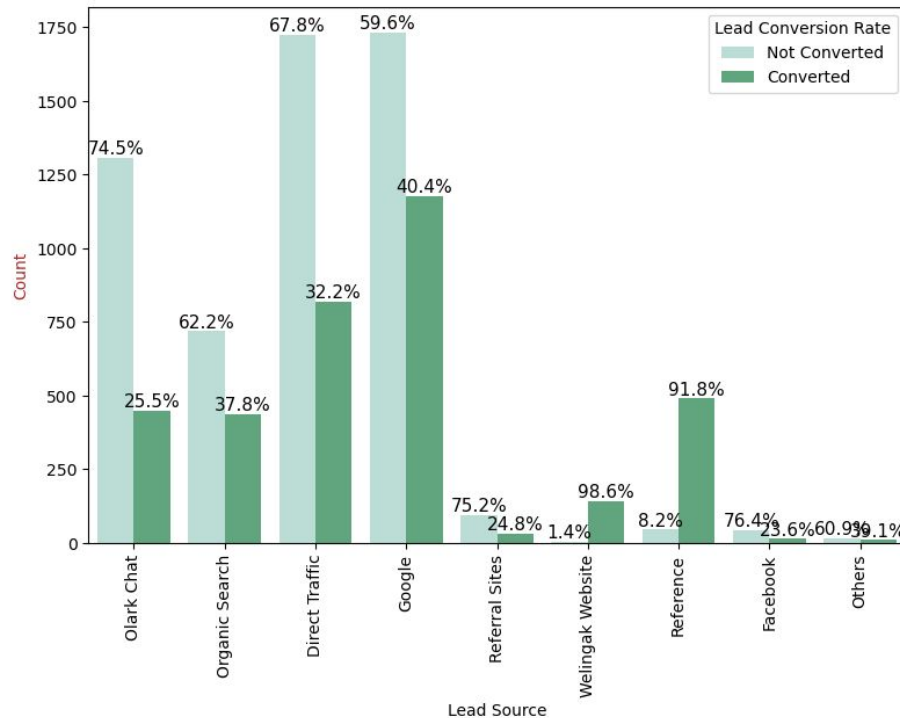
Business Goal

The company requires a model to be built for selecting the most promising leads.

Lead score to be given to each leads such that it indicates how promising the lead could be. Higher the lead score, more promising the lead to get converted.

The model to be built in lead conversion rate around 80% or more.
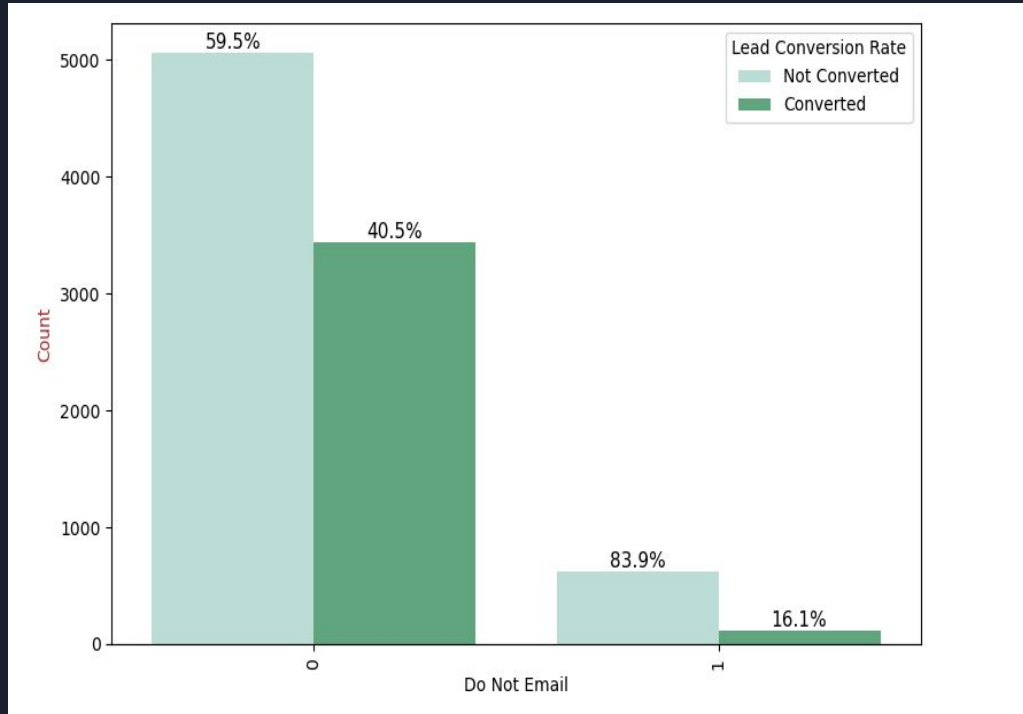
Exploratory Data Analysis

Lead Source Vs Converted

Google searches had high conversions as compared to other modes, while references had high conversion rates.
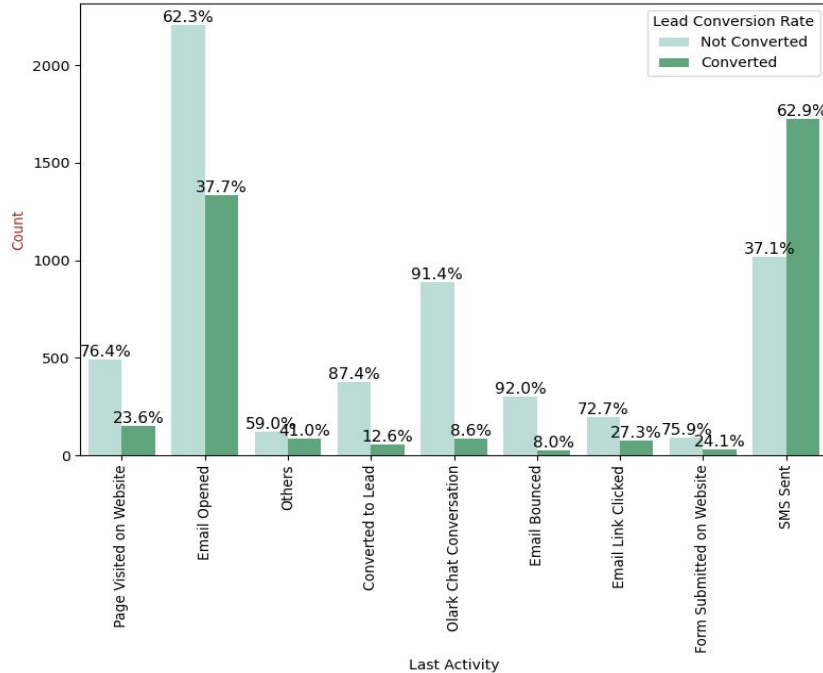
Exploratory Data Analysis

Do Not Email Vs Converted

Google searches had high conversions as compared to other modes, while references had high conversion rates.
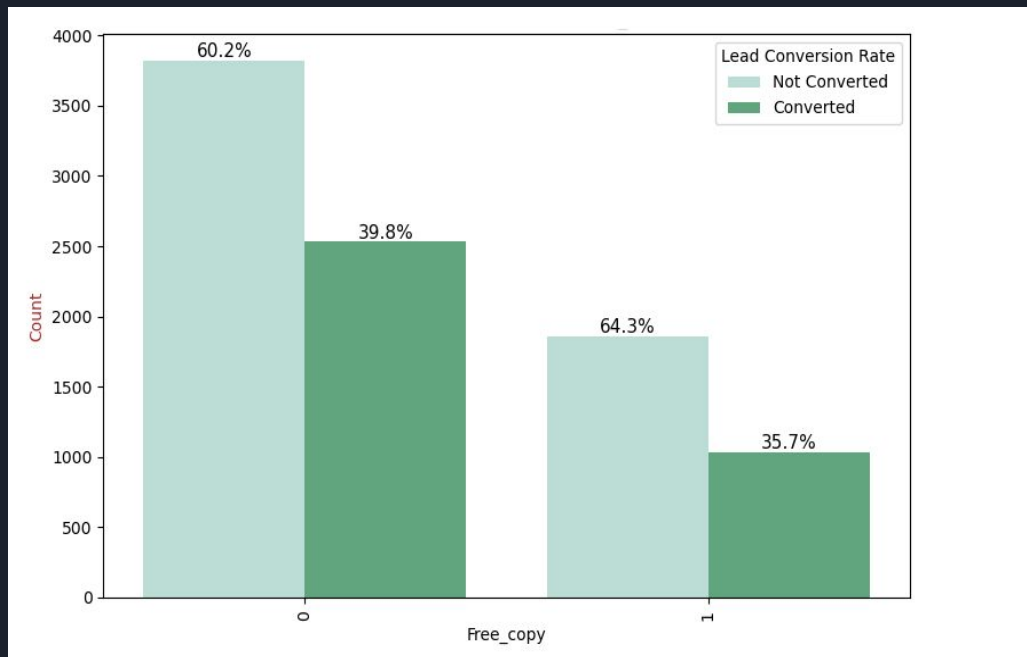
Exploratory Data Analysis

Last Activity Vs Converted

SMS has shown to be more promising method for having higher confirmed leads.
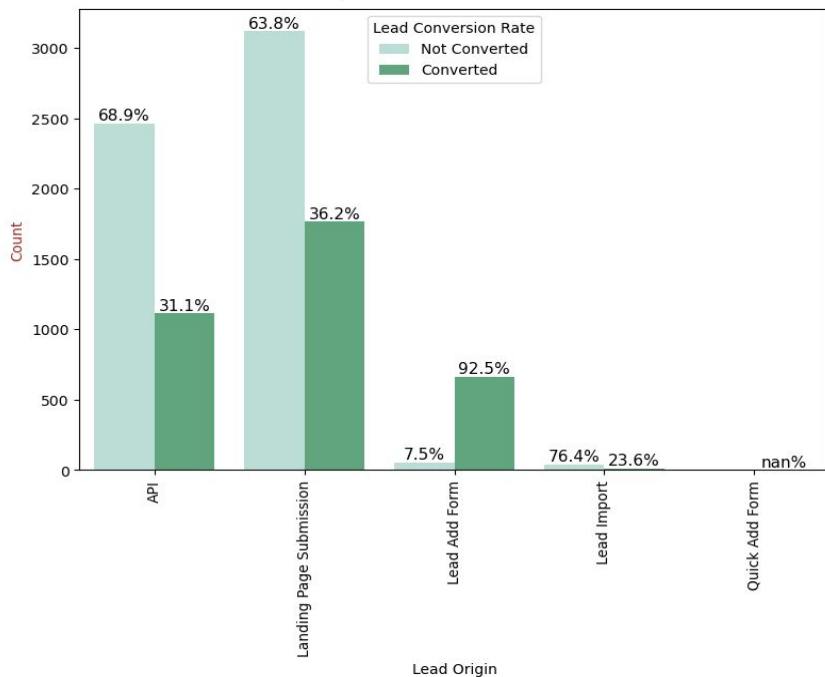
Exploratory Data Analysis

Free Copy Vs Converted

Copies of Interviews are preferred less by Leads.
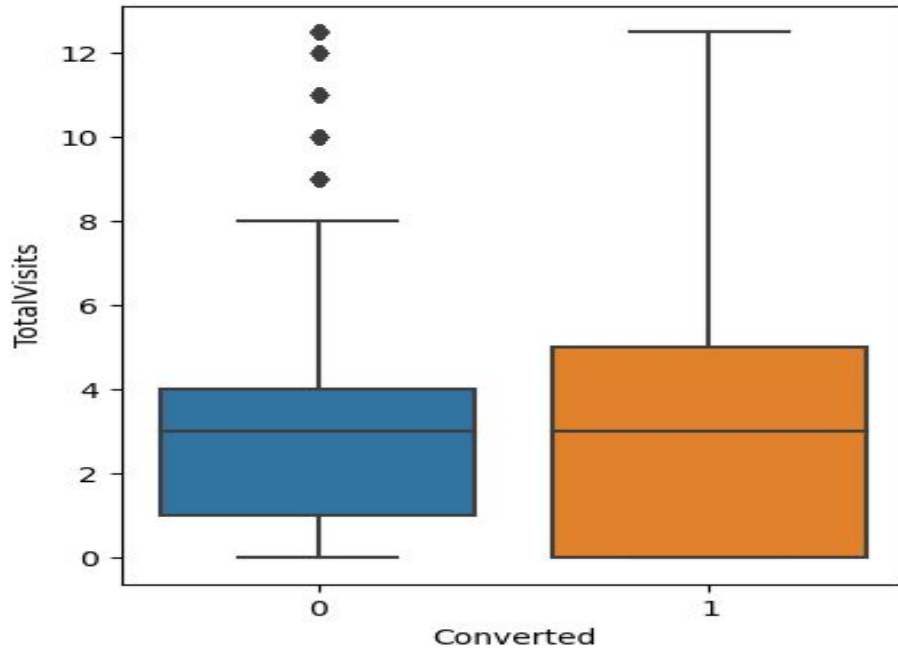
Exploratory Data Analysis

Lead Origin Vs Converted

Landing Page Submissions had high lead conversions.

Exploratory Data Analysis

Total Visits Vs Converted

Total Visits have a slighter higher chances of being a promising leads.

Exploratory Data Analysis
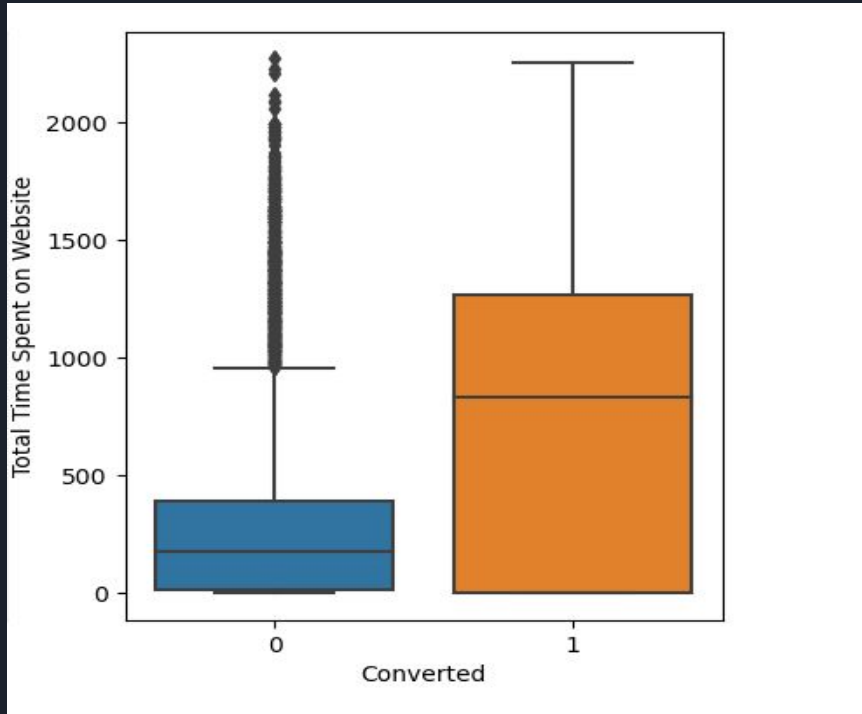
Total Time Spent on Website Vs Converted

People spending higher time than average time are more promising leads.

Model Building

➢ Splitting into train and test sets

➢ Scale variables in train set

➢ Build first model

➢ Use RFE to eliminate less relevant variables

➢ Build next model

➢ Eliminate variables by checking the high p-values

➢ Check VIF values for all variables

➢ Evaluate accuracy and other parameters

➢ Precision and Recall Analysis on test predictions

Model Evaluation


```
                Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6468
Model:                            GLM   Df Residuals:                     6455
Model Family:                Binomial   Df Model:                           12
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                 -2743.1
Date:                Mon, 18 Sep 2023   Deviance:                        5486.1
Time:                        03:39:03   Pearson chi2:                  8.11e+03
No. Iterations:                     7   Pseudo R-squ. (CS):             0.3819
Covariance Type:            nonrobust
==============================================================================
                                       coef    std err      z   P>|z|   [0.025    0.975]
------------------------------------------------------------------------------
const                                -1.0236    0.143   -7.145   0.000   -1.304   -0.743
Total Time Spent on Website           1.0498    0.039   27.234   0.000    0.974    1.125
Lead Origin_Landing Page Submission  -1.2590    0.125  -10.037   0.000   -1.505   -1.013
Lead Source_Olark Chat                0.9072    0.118    7.701   0.000    0.676    1.138
Lead Source_Reference                 2.9253    0.215   13.615   0.000    2.504    3.346
Lead Source_Welingak Website          5.3887    0.728    7.399   0.000    3.961    6.816
Last Activity_Email Opened            0.9421    0.104    9.022   0.000    0.737    1.147
Last Activity_Olark Chat Conversation -0.5556   0.187   -2.974   0.003   -0.922   -0.189
Last Activity_Others                  1.2531    0.238    5.259   0.000    0.786    1.720
Last Activity_SMS Sent                2.0519    0.107   19.106   0.000    1.841    2.262
Specialization_Hospitality Management -1.0944   0.323   -3.391   0.001   -1.727   -0.462
Specialization_Others                -1.2033    0.121   -9.950   0.000   -1.440   -0.966
Current_occupation_Working Professional 2.6697  0.190   14.034   0.000    2.297    3.042
==============================================================================
```
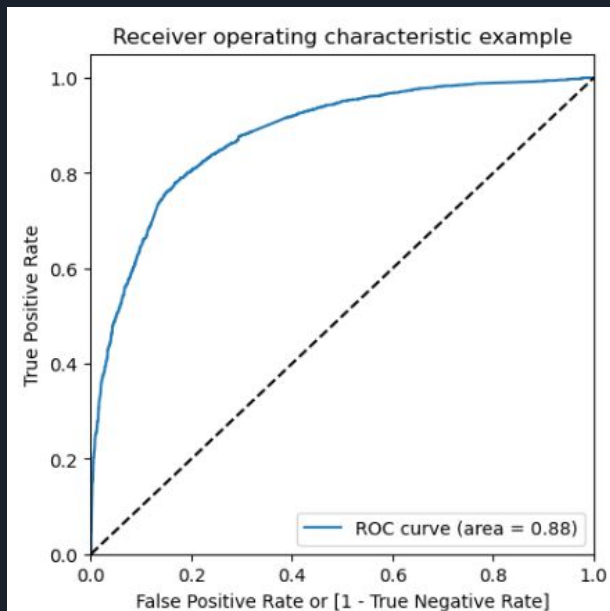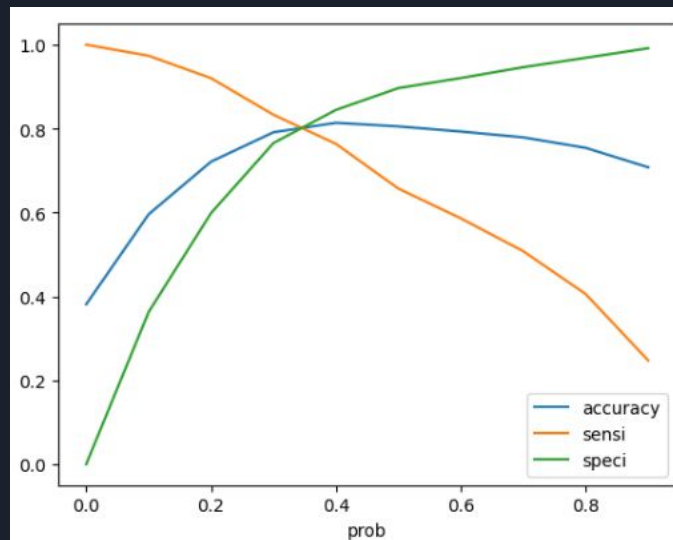
All p-values are 0.

ROC Curve

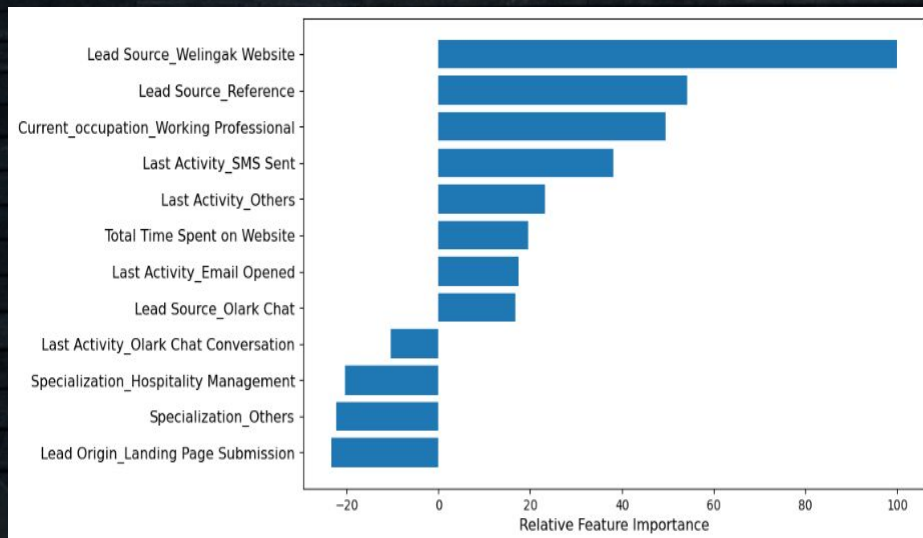Area under the curve = 0.88

Finding Optimal Threshold

Graph showing changes in sensitivity, specificity and accuracy with changes in the probability threshold values.
Optimal Threshold = 0.35

Feature Importance

Three variables which contribute the most in probability of a lead conversion are:

- Lead Source_Welingak Website
- Lead Source_Reference
- Current_occupation_Working Professional

These are dummy features created from categorical variables.

All three contribute positively towards the lead conversion.

These results indicates that the company should focus more on the leads with these variables.

Recommendations

- There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom.
- In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.
- First, sort out the best prospects from the leads you have generated. TotalVisits, Page Views Per Visit, Total Time Spent on Website, Lead Source_Welingak Website, Lead Source_Reference, Current_occupation_Working Professional which contribute most towards the probability of a lead getting converted.
- Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies.
- Monitor each lead carefully so that you can tailor the information you send to them.
- Carefully provide job offerings, information or courses that suits best according to the interest of the leads. A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.
- Focus on converted leads. Hold question-answer sessions with leads to extract the right information you need about them.
- Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.

THANK YOU