Lead Scoring Case Study-DS C50

- Sweta Baghel
- Tamilrasu

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%.

PROBLEM STATEMENT

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.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.

X Education wants to 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 higher lead score h have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

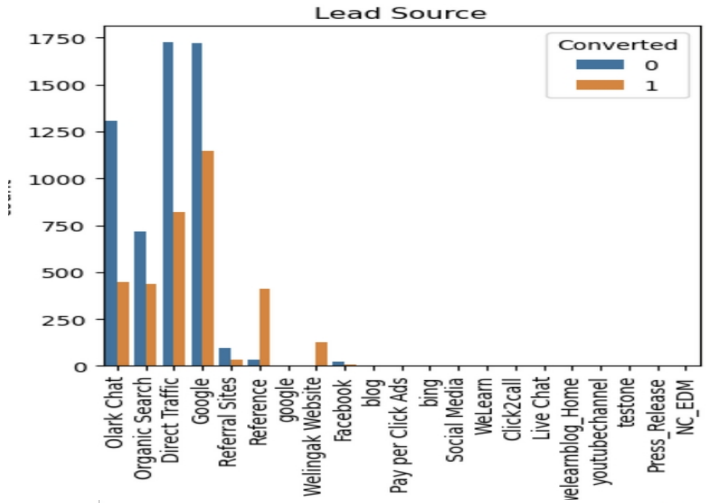
GOAL OF CASE STUDY

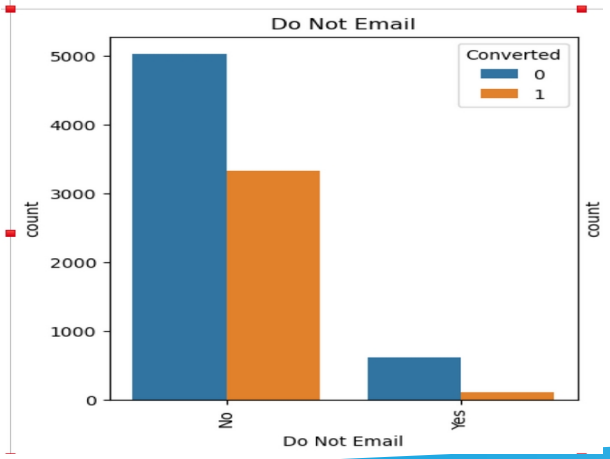
1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

STRATEGY:

- Import data
- Clean and prepare the acquired data for further analysis
- Exploratory data analysis for figuring out most helpful attributes for conversion
- Scaling features
- Prepare the data for model building
- Build a logistic regression model
- Assign a lead score for each leads
- Test the model on train set
- Evaluate model by different measures and metrics
- Test the model on test set
- Measure the accuracy of the model and other metrics for evaluation





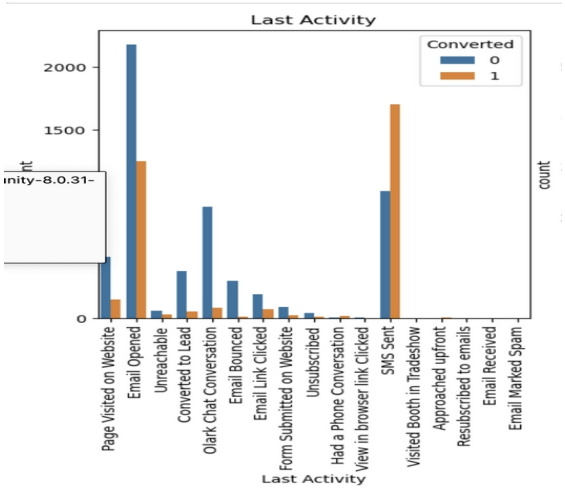
EXPLORATORY DATA ANALYSIS:

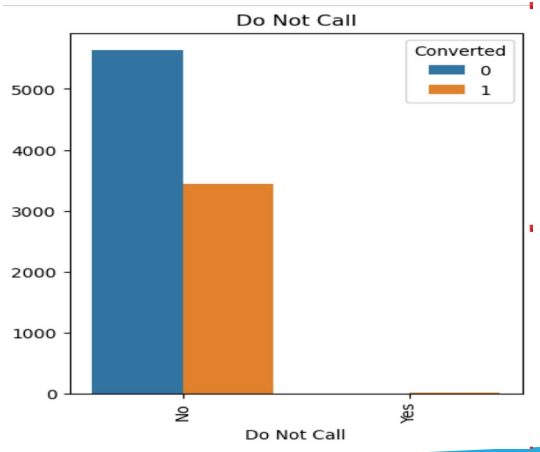
LEAD SOURCE VS CONVERTED:

google searches has had high conversions compared to other modes, whilst references has had high conversion rate.

DO NOT EMAIL VS CONVERTED:

google searches has had high conversions compared to other modes, whilst references has had high conversion rate.



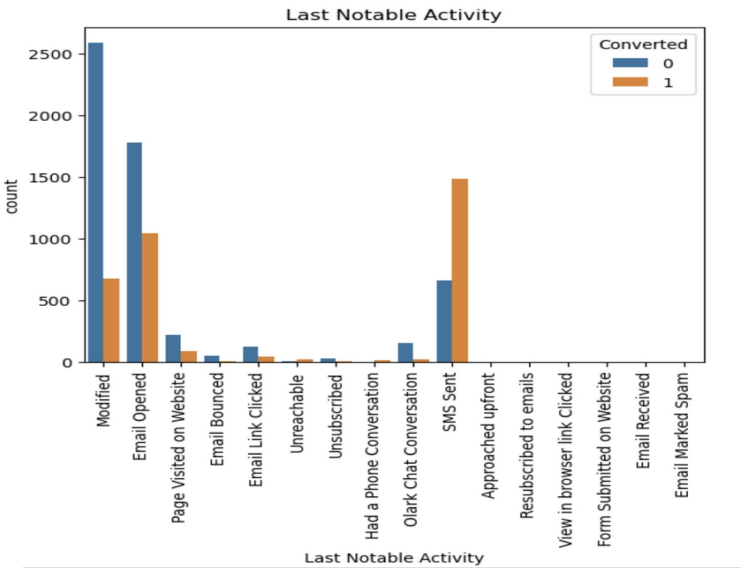


LAST ACTIVITY VS CONVERTED:

SMS has shown to be a promising method for getting higher confirmed leads, emails also has high conversions.

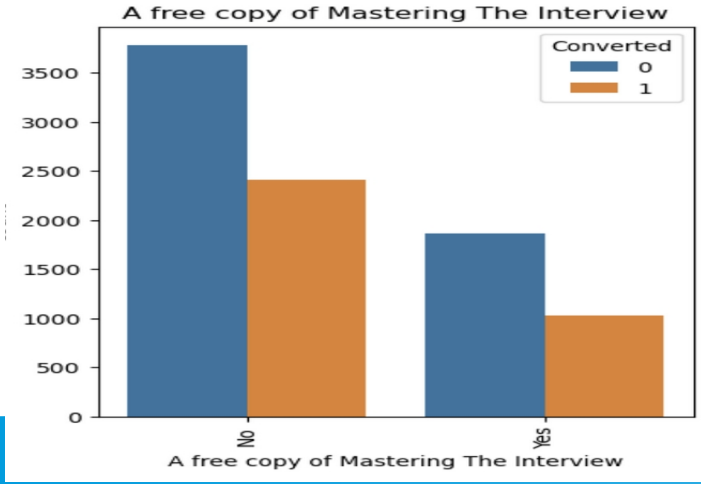
DO NOT CALL VS CONVERTED:

most leads prefer not to informed through phone



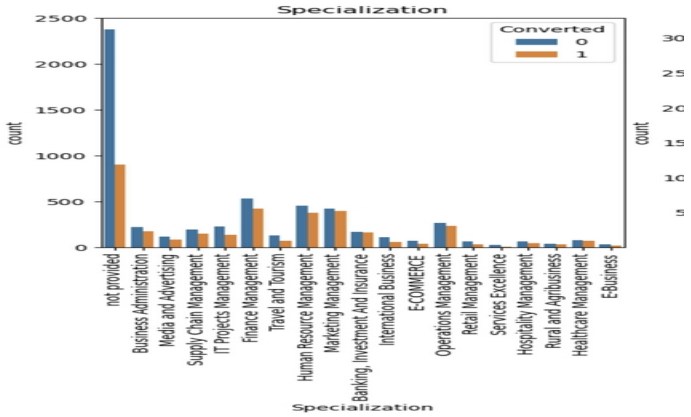
LAST NOTABLE ACTIVITY VS CONVERTED:

Most leads are converted with messages. Emails also induce leads.



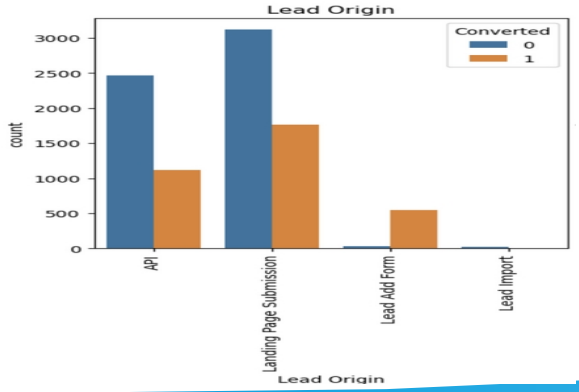
FREE COPY OF MASTERING THE INTERVEIW VS CONVERTED:

Leads prefer less copies of interviews.



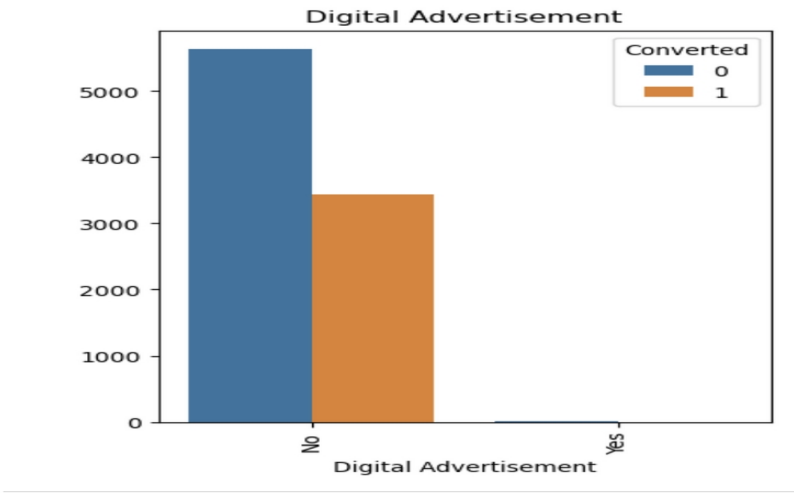
SPECIALIZATION VS CONVERTED:

Most of the leads have no information about specialization. On the other hand,marketing management, human resources management has high conversion rates. people from these specializations can be promising leads



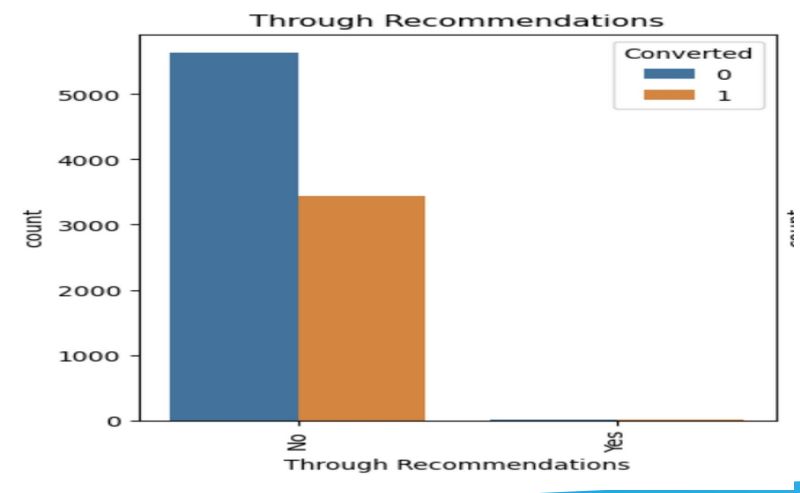
LEAD ORIGIN VS CONVERTED:

Landing page submissions has had high lead conversions



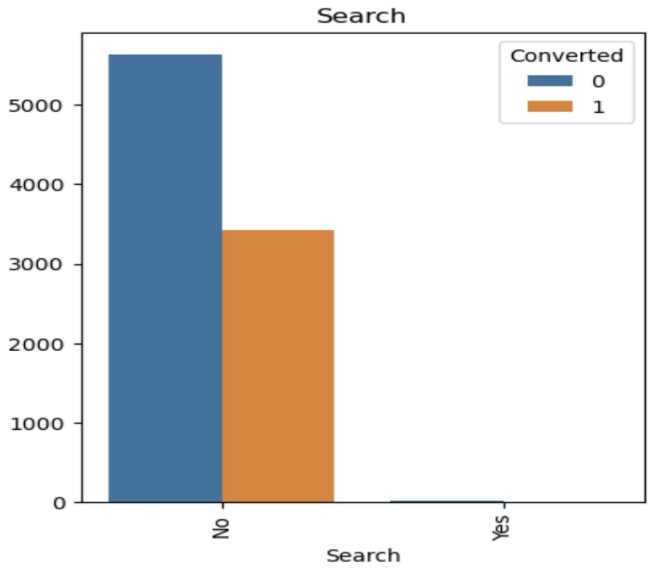
DIGITAL ADVERTISEMENTS VS CONVERTED:

Based on the above graph digital advertisements do not have promising leads



THROUGH RECOMMENDATIONS VS CONVERTED:

From the above graph, recommendations are not a good source for promising leads

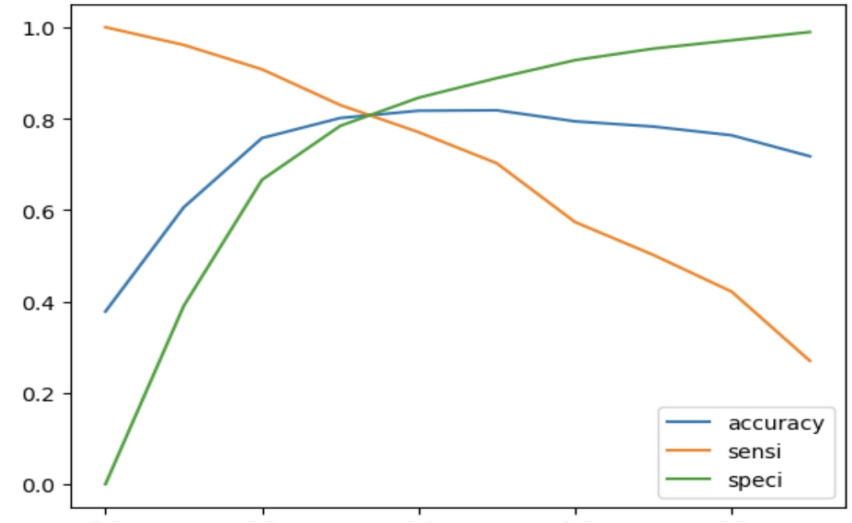


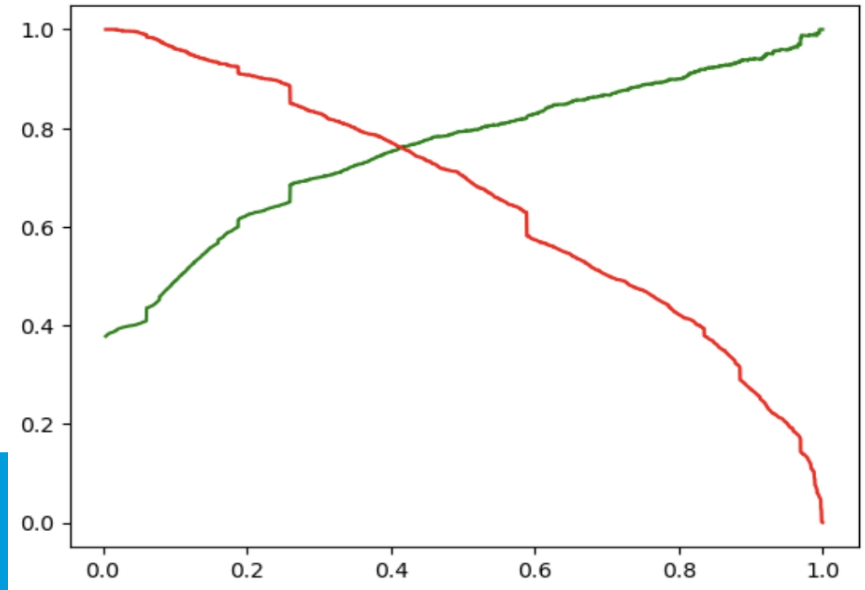
SEARCH VS CONVERTED:

Graph shows searches are not good source of leads

MODEL BUILDING:

- Splitting into train and test set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high p-values
- Check VIF value for all the existing columns
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test predictions





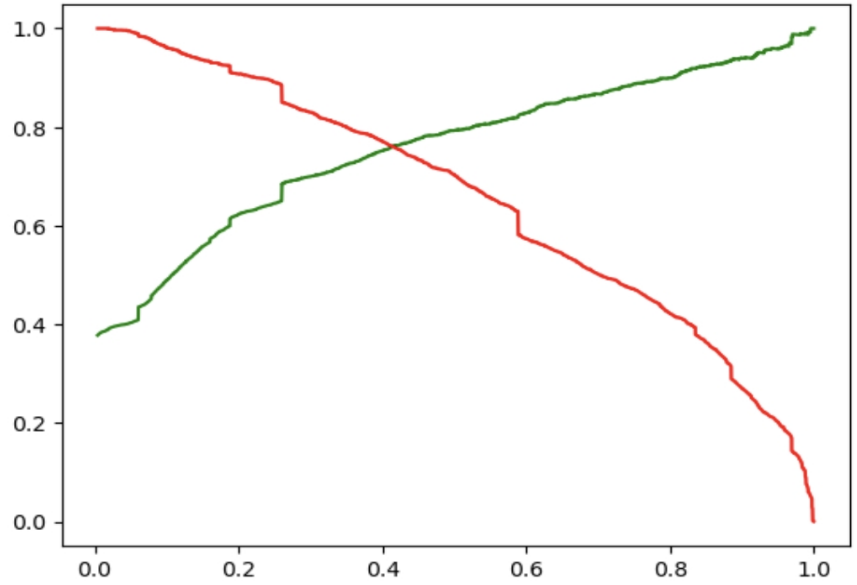
MODEL EVALUATION (TRAIN):

ACCURACY SENSITIVITY AND SPECIFICITY

- 80.9% Accuracy
- 77.6% Sensitivity
- 82.9% Specificity

PRECISION AND RECALL

- 73.4% Precision
- 77.6% Recall



PRECISION AND RECALL

- 74.4% Precision
- 75.5% Recall

Test set threshold has been set as 0.41

ACCURACY SENSITIVITY AND SPECIFICITY

- 80.1% Accuracy
- 75.5% Sensitivity
- 83.1% Specificity

SUMMARY:

EDA:

- People spending higher than average time are promising leads, so targeting them and approaching them can be helpful in conversions
- SMS messages can have a high impact on lead conversion
- Landing page submissions can help find out more leads
- Marketing management, human resources management has high conversion rates. People from these specializations can be promising leads
- References and offers for referring a lead can be good source for higher conversions
- An alert messages or information has seen to have high lead conversion rate

Logistic Regression Model:

- The model shows high close to 81% accuracy
- The threshold has been selected from Accuracy, Sensitivity, specificity measures and precision, recall curves.
- The model shows 76% sensitivity and 83% specificity
- The model finds correct promising leads and leads that have less chances of getting converted
- Overall this model proves to be accurate