LEAD SCORING ASSIGMENT

Submitted by :- Sushmita Roy, Chetna Turankar and Utteya Pal

PROBLEM STATEMENT

- X Education is a company that sells classes online to people who work in certain industries. They advertise their classes on various websites and search engines like Google. When someone visits their website, they might look at the classes, sign up for one by filling out a form, or watch some videos.
- Now, when people fill out a form and provide their email or phone number, the company considers them as potential customers, or "leads." The company also gets leads from people who were referred to them by others in the past.
- Once they have these leads, the sales team at the company starts reaching out to them—making phone calls, sending emails, and so on. Through this process, some of the leads decide to sign up for a class, but many of them don't. At X Education, about 30% of these leads end up becoming actual customers.

STRATEGY

First, we need to gather the information for our study. Once we have the data, we'll clean it up and get it ready for analysis. We'll explore the data to understand it better. After that, we'll scale the features to make sure they're all in a similar range.

Next, we'll split the data into two parts: one for testing and one for training our model. Then, we'll create a logistic regression model to predict something called a 'Lead Score.' This score helps us understand how likely someone is to become a customer.

To see how well our model works, we'll use different measurements like specificity and sensitivity, or precision and recall. Finally, we'll apply our best model to the test data based on these measurements to make sure it performs well."

PROBLEM SOLVING METHODOLOGY

1. Getting and Preparing Data

- Get the data from where it's stored.
- Make sure the data is in a clean format that we can easily analyze.
- Remove any duplicate information.
- Treat any unusual or extreme data points.
- Explore the data to understand it better.
- Standardize the features to make them consistent and comparable.

2. Getting Ready for Analysis:

Make sure all the numbers in our data are on a similar scale (Feature Scaling).

Divide our data into two parts: one for training our analysis and the other for testing it out.

3. Creating the Model:

Choose the most important features using a method called RFE (Feature Selection).

Find the best model using Logistic Regression.

Measure how well the model works by looking at different things like overall accuracy, sensitivity (catching the positives), specificity (avoiding false alarms), precision (correctly identifying positives), and recall (catching all the actual positives)

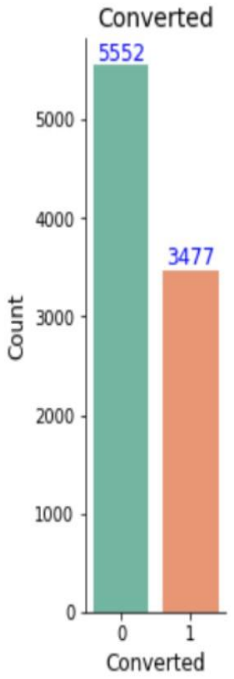
4. Outcome:

Figure out the lead score and see if our final predictions show an 80% chance of turning potential customers into actual buyers.

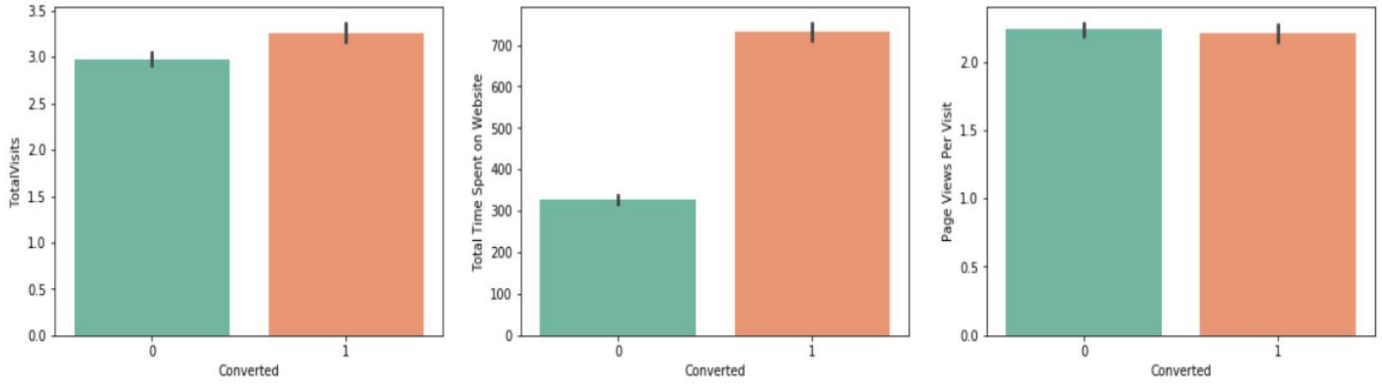
Test our final predictions using a specific measure we get from looking at sensitivity and specificity metrics.

EDA

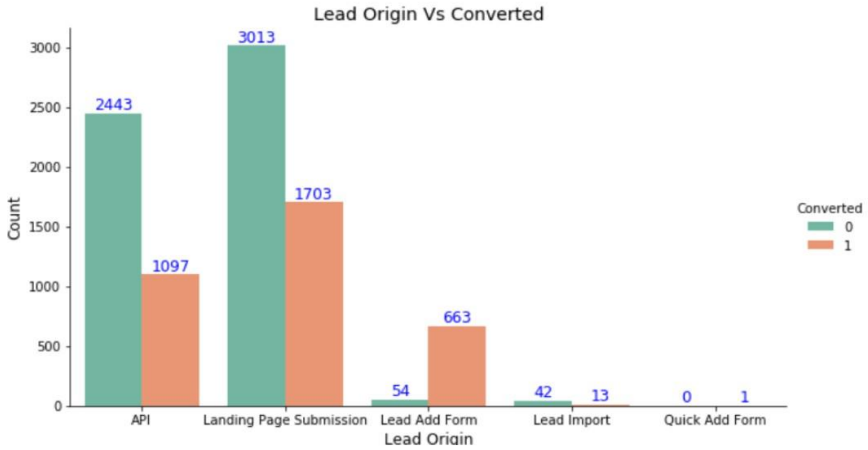
We have around 39% Conversion rate in Total



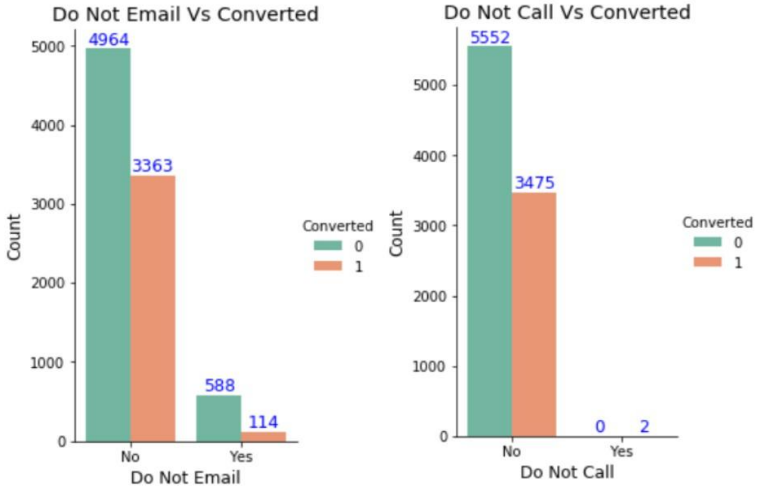
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit



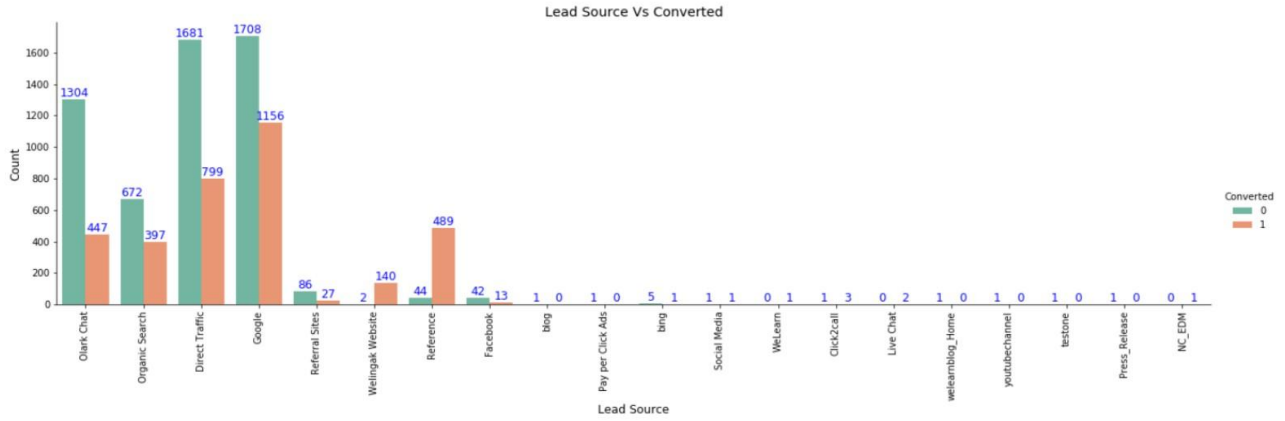
In Lead Origin, maximum conversion happened from Landing Page Submission



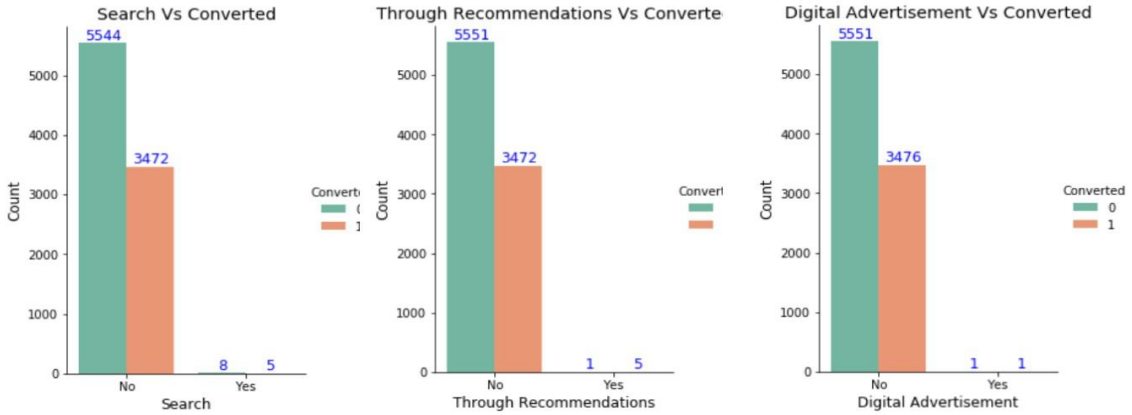
Major conversion has happened from Emails sent and Calls made



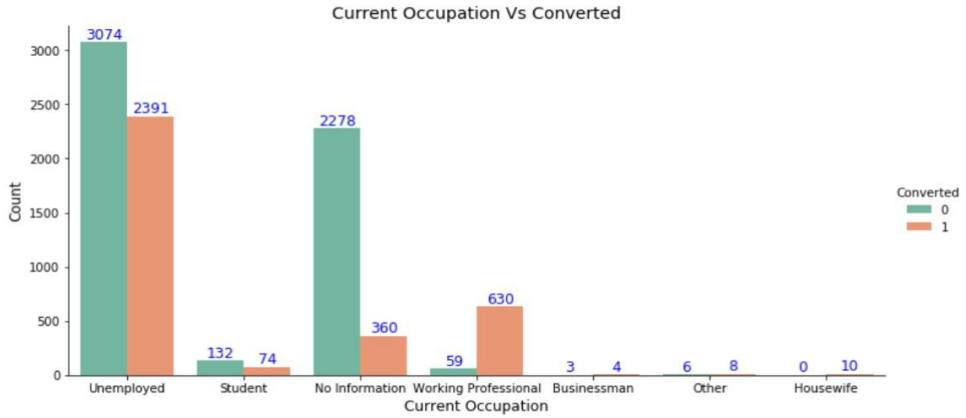
Major conversion in the lead source is from Google



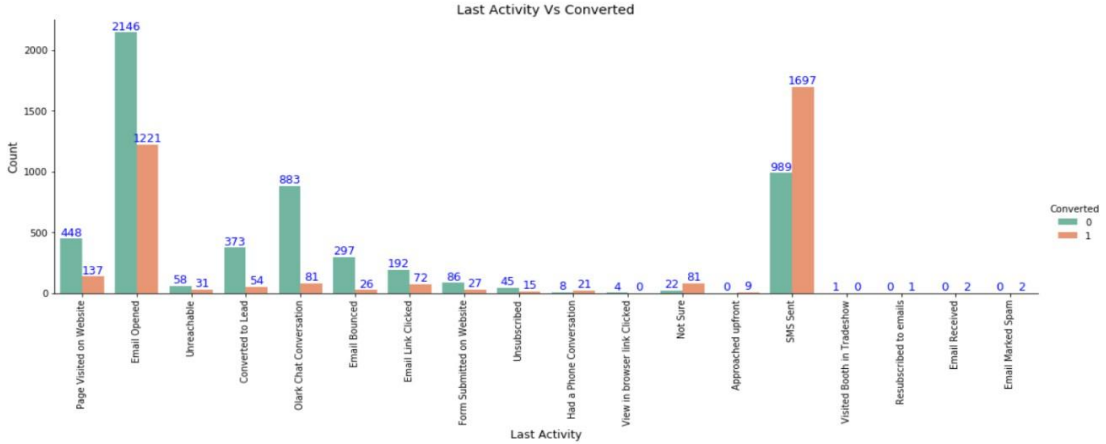
Not much impact on conversion rates through Search, digital advertisements and through recommendations



More conversion happened with people who are unemployed



Last Activity value of SMS Sent' had more conversion.

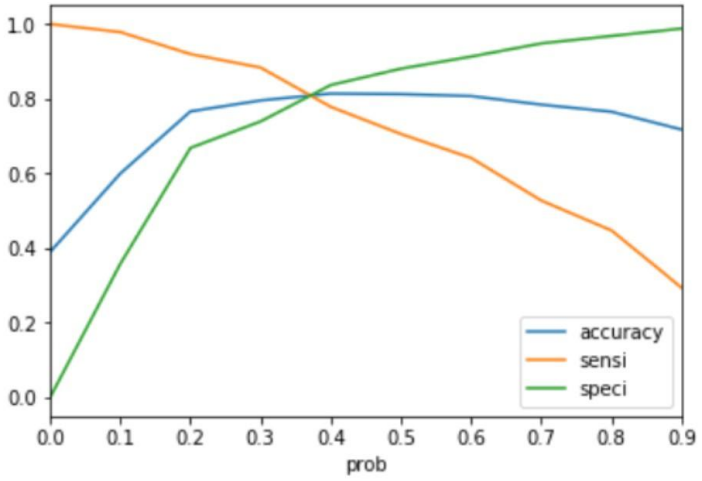


FACTORS AFFECTING THE LIKELIHOOD OF PEOPLE BUYING:

- Whether or not they want to receive emails.
- How many times they've visited the website.
- The total time they've spent on the website.
- How they initially showed interest, like submitting a form or using a chat.
- Where they found out about the service, like through chat or the website.
- Their recent interactions, like bounced emails or SMS communication.
- Their current job status, whether they're working or not.
- Notable recent activities, such as having a phone conversation or being unreachable.

MODEL EVALUATION- SENSITIVITY AND SPECIFICITY

The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



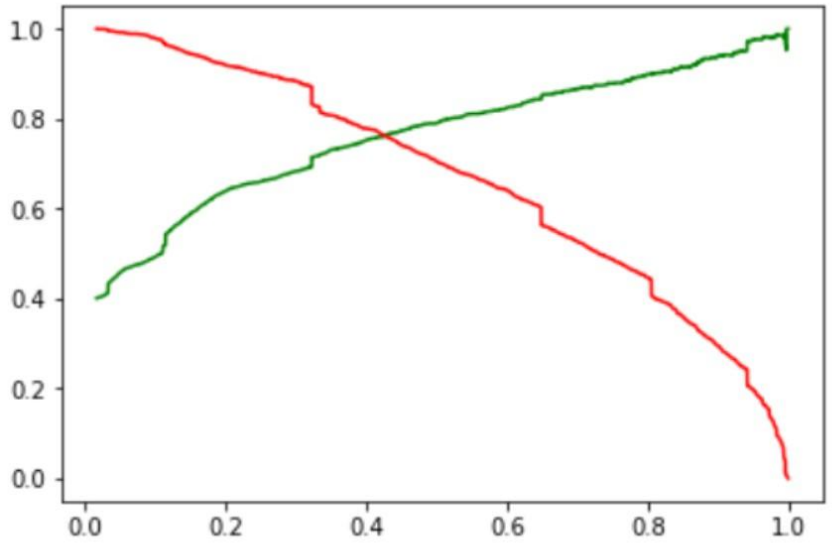
Confusion Matrix

3161	697
974	1965

- Accuracy - 81%
- Sensitivity - 80 %
- Specificity - 82 %
- False Positive Rate - 18 %
- Positive Predictive Value - 74 %
- Positive Predictive Value – 86%

MODEL EVALUATION- PRECISION AND RECALL

The graph depicts an optimal cut off of 0.42 based on Precision and Recall



Confusion Matrix

3397	461
725	1737

- Precision - 79 %
- Recall - 71 %

CONCLUSION

- We looked at different ways to measure how well our predictions work, focusing on sensitivity and specificity. We decided on the best cut-off point based on these measures.
- The accuracy, sensitivity, and specificity of our predictions on the test set are all pretty good, around 81%, 79%, and 82% respectively. These values are close to what we saw during the training.
- The lead score we calculated suggests that our model predicts a conversion rate of about 80% for the training set and 79% for the test set.
- The top three factors that seem to influence whether a lead becomes a customer are the total time spent on the website, using the Lead Add Form, and having a phone conversation based on the last notable activity.
- So, overall, it looks like our model is doing well.