LEAD SCORING CASE STUDY

PRESENTED BY:

DHRUV SINHA

SUDEEP MENON

PROBLEM STATEMENT:

- ✓ An education company named X Education sells online courses to industry professionals.
- ✓ While X Education gets a lot of leads, its lead conversion rate is a mere 30%.
- ✓ 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.

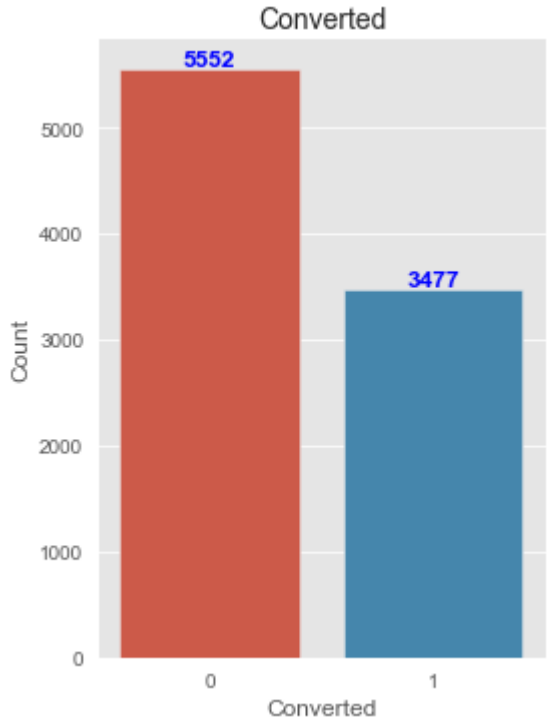
BUSINESS OBJECTIVE:

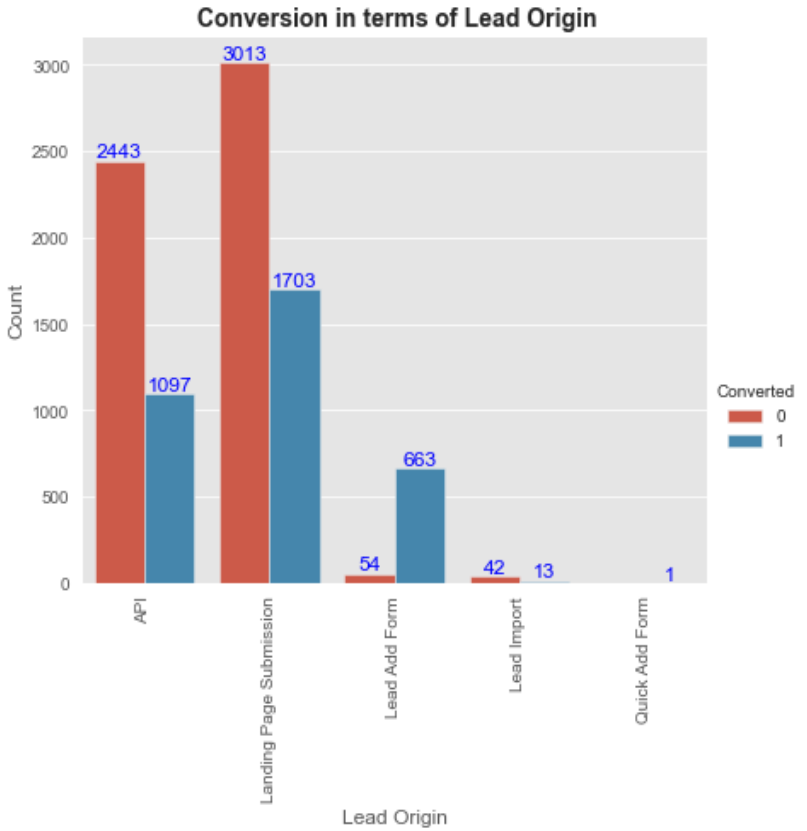
- ✓ Build a logistic regression model to identify hot leads with a ballpark of the target lead conversion rate of ~80%
- ✓ Model should be flexible and should be able to incorporate company's future requirements.

OVERALL APPROACH:

- ✓ READING AND UNDERSTANDING THE DATA
- ✓ DATA CLEANING AND MANIPULATION
- ✓ EXPLORATORY DATA ANALYSIS
- ✓ TEST TRAIN SPLIT
- ✓ DUMMY VARIABLE CREATION AND FEATURE SCALING
- ✓ BUILDING A LOGISTIC REGRESSION MODEL
- ✓ MODEL EVALUATION: SENSITIVITY, SPECIFICITY AND PRECISION
- ✓ PREDICTIONS ON THE TEST SET
- ✓ INSIGHTS AND RECOMMENDATIONS

EXPLORATORY DATA ANALYSIS:



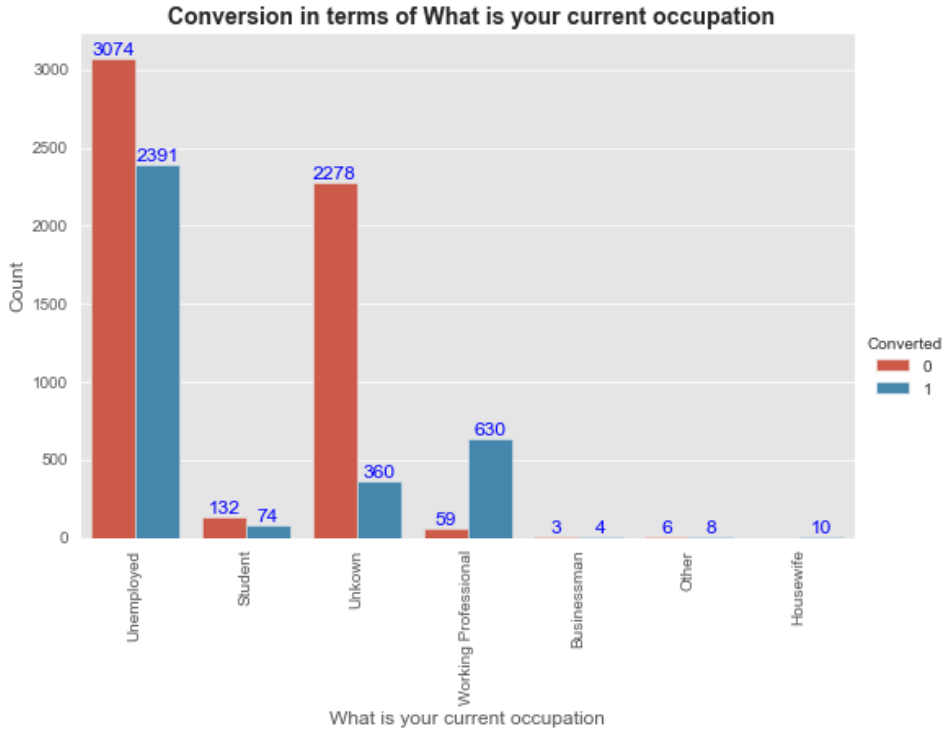


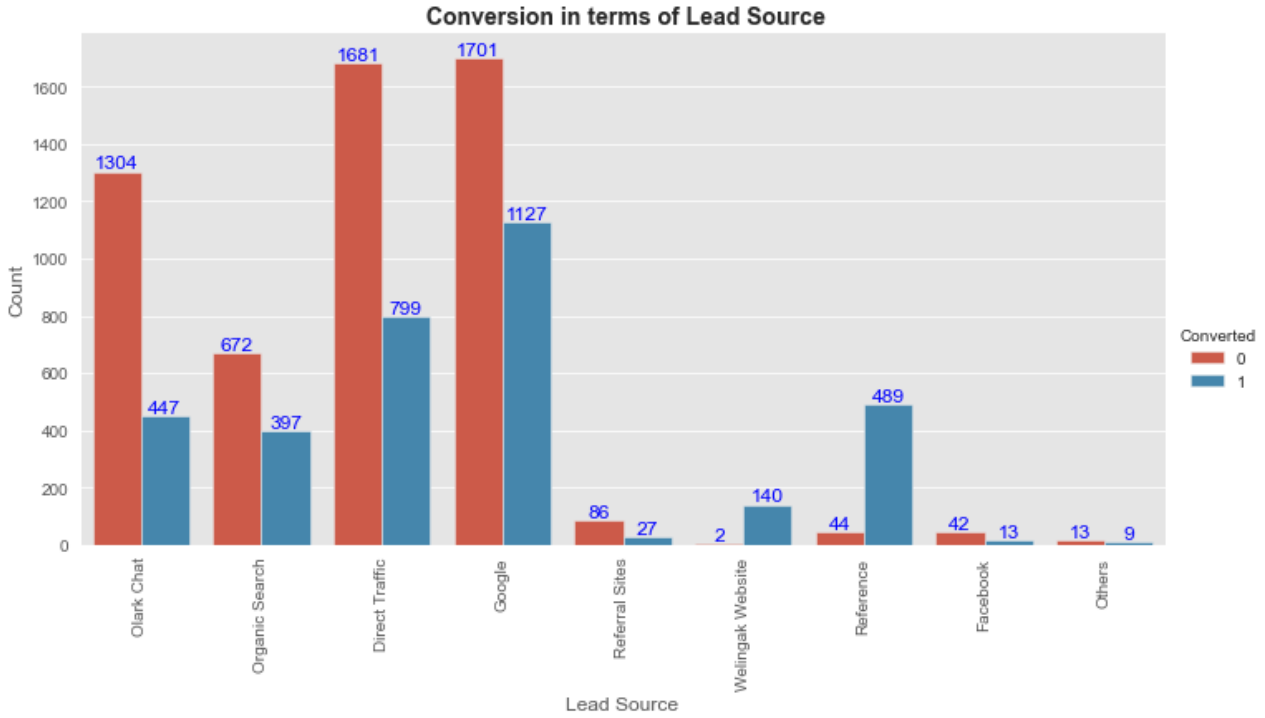
Lead conversion ratio for X Education is ~39%

LEAD ORIGIN:

- Leads from 'API' and 'Landing Page Submission' are high; conversion rates are poor at 31% and 36% respectively.
- Though leads from 'Lead Add Form' is low, the conversion rate is very high at ~92%.

EXPLORATORY DATA ANALYSIS:





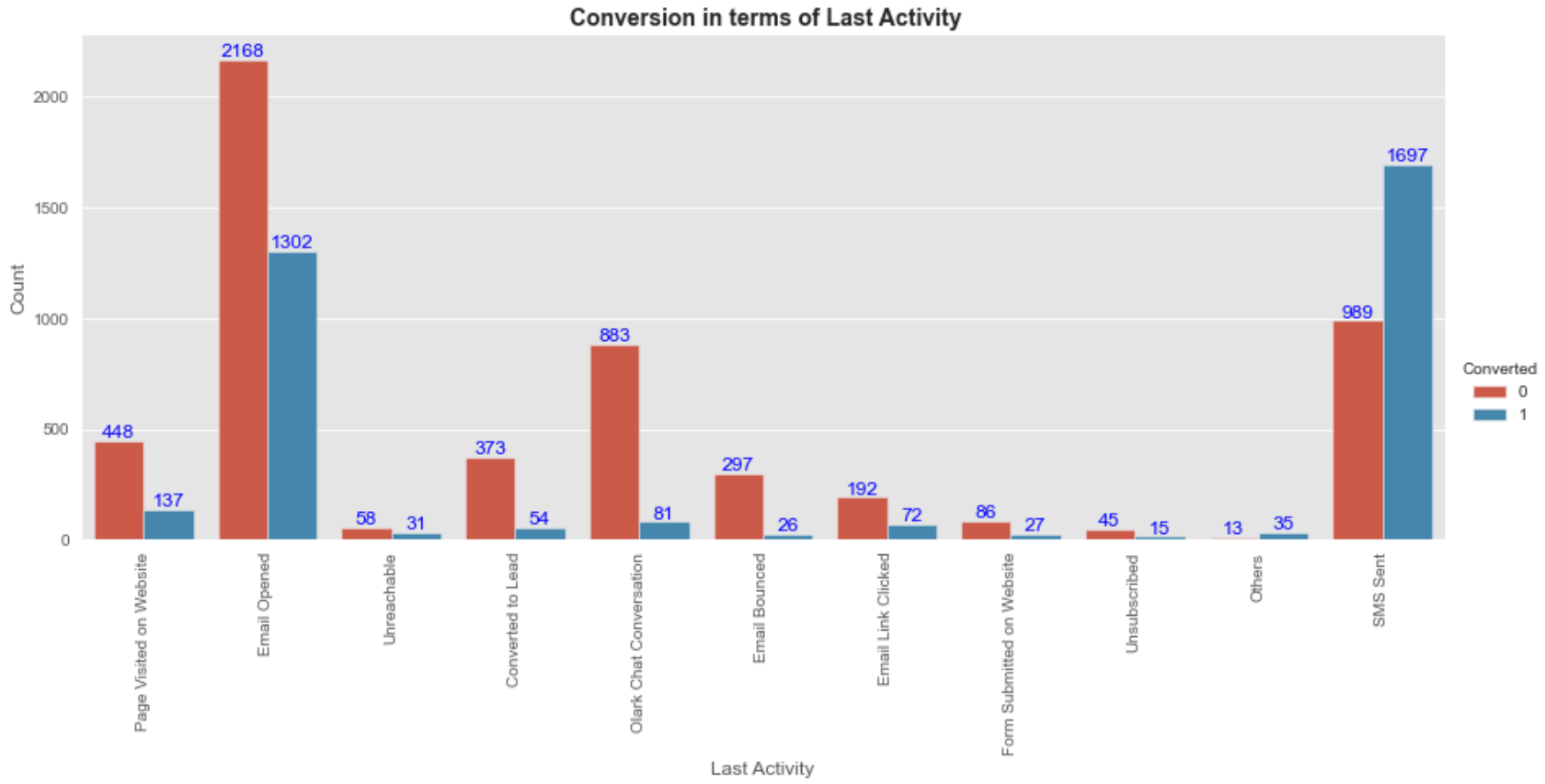
WHAT IS YOUR CURRENT OCCUPATION:

- Working Professionals have the highest conversion rate at ~91% followed by students at ~86%.
- Unemployed leads have low conversion rates but they generate maximum leads counts.

LEAD SOURCE:

- ~90% of the leads are generated from 4 Lead Sources 'Google','Direct Traffic', 'Olark Chat' and 'Organic Search’.
- Only 34% of those leads are being converted.
- The conversion rate for 'Welingak Website' and 'Reference' are high at ~99% and ~92% respectively, however only 6% of the total leads are being generated from these sources.

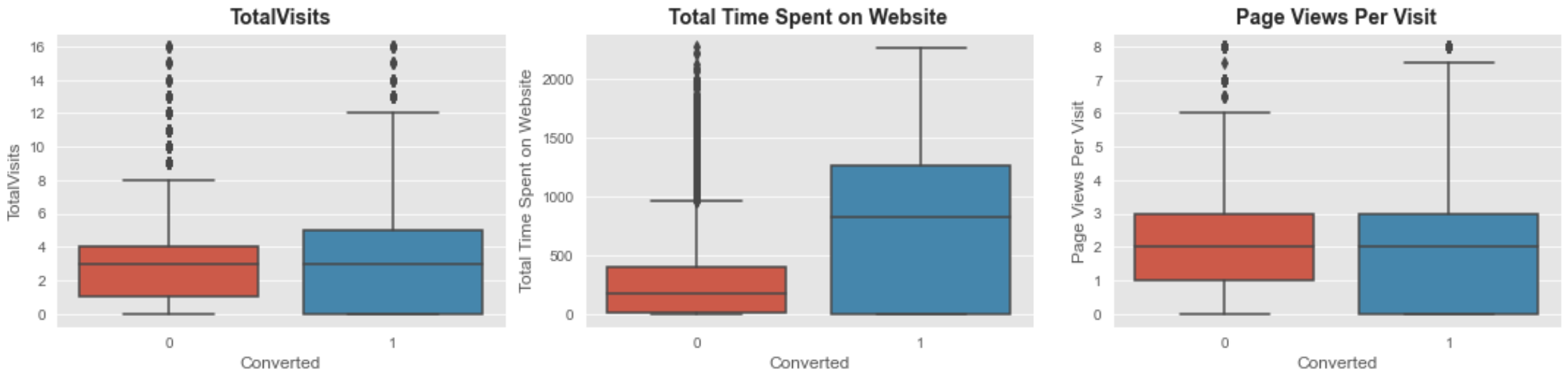
EXPLORATORY DATA ANALYSIS:



LAST ACTIVITY:

- Maximum leads are generated from people with last activity - Email opened and SMS sent.
- Conversion rate is highest for SMS Sent (~63%), where as it is only ~38% for Email Opened.
- Olark chat conversation and Page Visited on Website generates significant number of leads but their conversion rates are extremely low at 8% and 38% respectively.

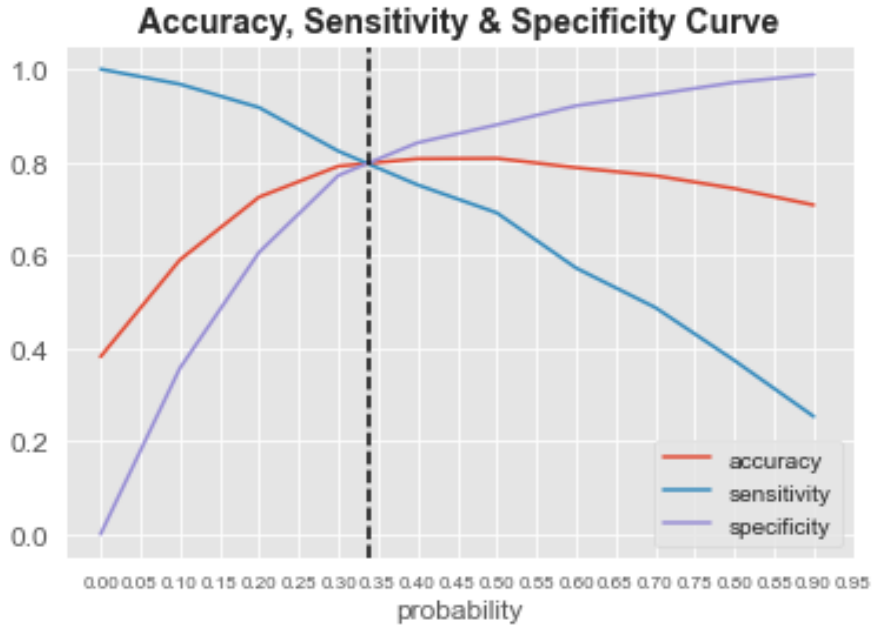
EXPLORATORY DATA ANALYSIS:



- **Total Visits:**
  - Median for converted and non-converted leads are same.
  - People who visits the platform have equal chances(50-50) of applying and not applying for the course.
- **Total Time Spent on Website:**
  - People spending more time on website have more chances of opting for a course
  - People who spend less time on the website didn't opt for any courses.
- **Page Views Per Visit:**
  - Median for converted and non-converted leads are same.

FINDING THE OPTIMAL CUT-OFF POINT:

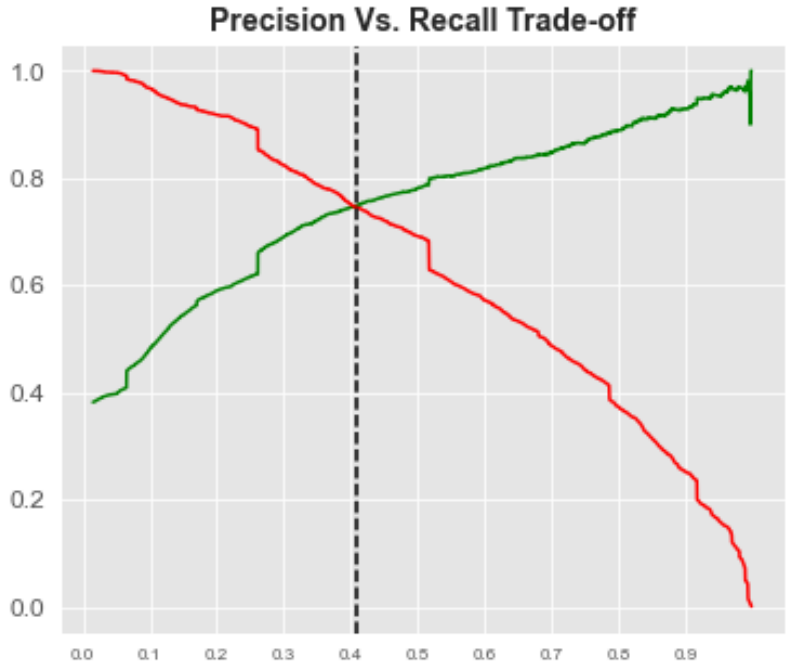
Final model cut-off considered the probability threshold of 0.34 based on model performance on Train Dataset



The above graph depicts the optimal probability cut-off of 0.34 based on accuracy, specificity and sensitivity analysis.

Model performance at 0.34 % cutoff:

Model Accuracy	:	80.1 %
Model Sensitivity	:	79.6 %
Model Specificity	:	80.4 %
Model Precision	:	71.5 %



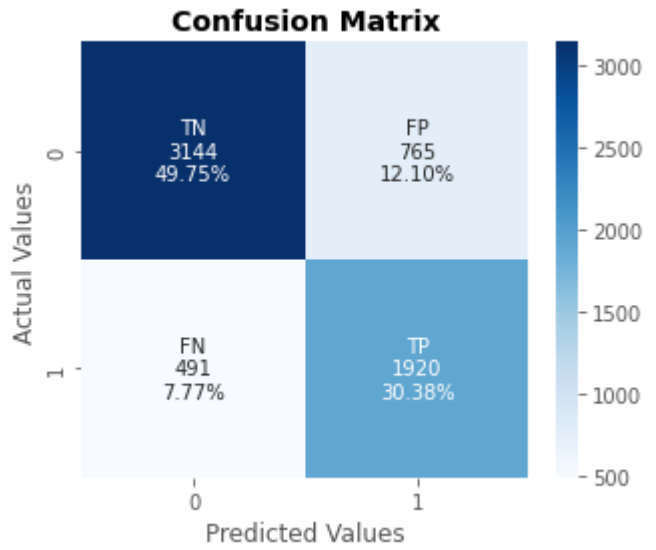
The graph depicts an optimal probability cut-off of 0.41 based on precision and recall trade-off analysis.

Model performance at 0.41 % cutoff:

Model Accuracy	:	80.8 %
Model Sensitivity	:	74.5 %
Model Specificity	:	84.7 %
Model Precision	:	75.0 %

MODEL EVALUATION:

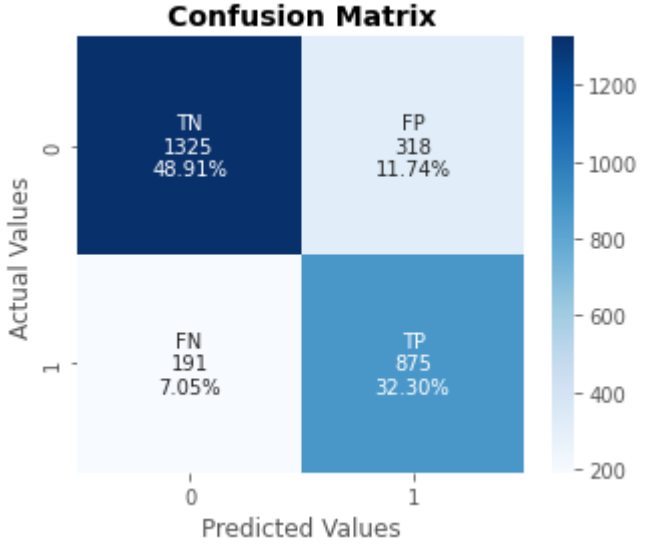
TRAIN DATA:



Model Accuracy : 80.1 %
Model Sensitivity : 79.6 %
Model Specificity : 80.4 %
Model Precision : 71.5 %

	precision	recall	f1-score	support
0	0.86	0.80	0.83	3909
1	0.72	0.80	0.75	2411
accuracy			0.80	6320
macro avg	0.79	0.80	0.79	6320
weighted avg	0.81	0.80	0.80	6320

TEST DATA:



Model Accuracy : 81.2 %
Model Sensitivity : 82.1 %
Model Specificity : 80.6 %
Model Precision : 73.3 %

	precision	recall	f1-score	support
0	0.87	0.81	0.84	1643
1	0.73	0.82	0.77	1066
accuracy			0.81	2709
macro avg	0.80	0.81	0.81	2709
weighted avg	0.82	0.81	0.81	2709

- ✓ Given the problem statement, focus should primarily should be on Sensitivity, Specificity and Precision.
- ✓ High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted.
- ✓ Whereas high Specificity will ensure that leads with border line probability of getting converted/not converted are not selected.
- ✓ On Train data, with probability cutoff = 0.34, the model performance parameters Sensitivity & Specificity are above ~80% , while Precision is ~72%.
- ✓ There is only ~2% difference on train and test data's performance metrics, implying that the final model didn't overfit training data and is performing well.
- ✓ Depending on the business requirement, we can increase or decrease the probability threshold value which in turn will decrease or increase the Sensitivity/Specificity of the model. (as explained in the next couple of slides)

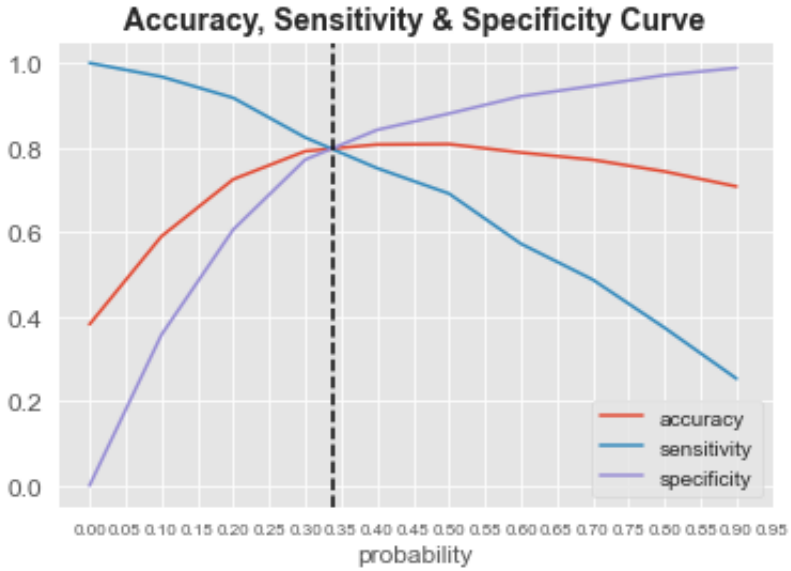
MODEL FLEXIBILITY:

Situation 1:

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.

Solution:

- Here, the concept of sensitivity is at play.
- $\boldsymbol{Sensitivity = True\ Positives / (True\ Positives + False\ Negatives)}$
- Sensitivity decreases with every increase in the cut-off threshold
- In the given situation, we will need a tweak the model to increase its sensitivity because high sensitivity will mean that our model will correctly predict almost all leads who are likely to convert.
- To achieve high sensitivity, we need to choose a low probability threshold value.
- However, sensitivity and precision are inversely correlated.
- In our case it would lead to our model misclassifying some of the non-converted leads as converted.



MODEL FLEXIBILITY:

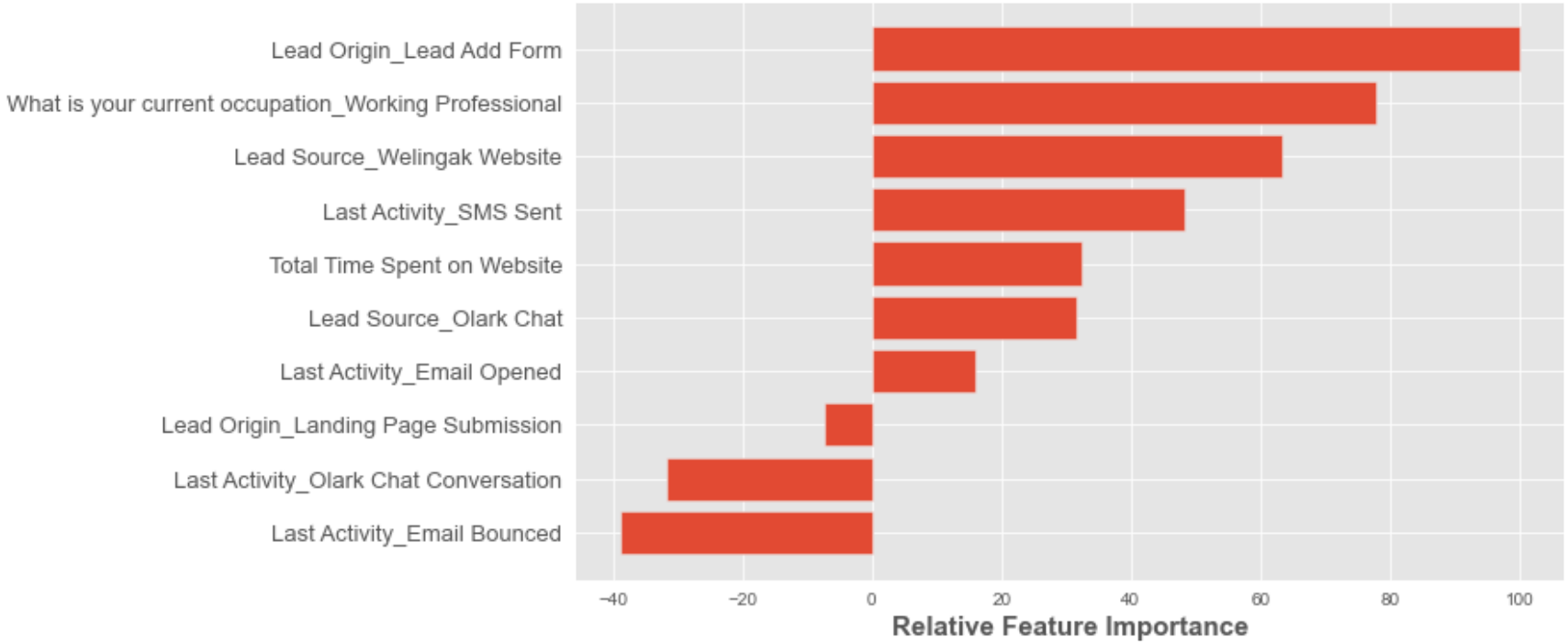
Situation 2:

At times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Solution:

- Here, the concept of specificity is at play.
- $\boldsymbol{Specificity = True\ Negatives / (True\ Negatives + False\ Positives)}$
- Specificity increases as the probability threshold increases.
- In the given situation, will need a high specificity because high specificity will mean that our model will correctly predict almost all leads who are not likely to convert. To achieve high specificity, we need to choose a high threshold value.
- However, increasing the specificity may lead to misclassifying some of the converted leads as non-converted.
- As the company has already reached its target for a quarter and doesn't want to make phone calls unless it's extremely necessary, it is a good strategy to go for high specificity.

VARIABLES IMPACTING THE CONVERSION RATE AND THE MODEL:



It was found that the variables that mattered the most in identifying potential customers are:

- ✓ Lead Origin: a. Lead Add Form b. Landing Page Submission
- ✓ Current occupation: a. Working Professional
- ✓ Lead source: a. Welingak website b. Olark Chat
- ✓ The Total Time Spent on the Website.
- ✓ Last activity was: a. SMS sent b. Email Opened C. Email Bounced
- ✓ Total time spent on the website

INSIGHTS AND RECOMMENDATIONS:

- ✓ Based our model, 7 of the 10 predictors belong to the below three variables:
  - ➢ Lead Origin: Lead add Form & Landing Page Submission
  - ➢ Lead Source: Welingak Website & Olark Chat.
  - ➢ Last Activity: SMS Sent, Email Opened, Email Bounced

- ✓ Based on the coefficient values in our model, the following are the top three categorical/dummy variables that should be focused on the most in order to increase the probability of lead conversion:
  - ➢ Lead Origin_Lead Add From
  - ➢ What is your current occupation_Working Professional
  - ➢ Lead Source_Welingak Website

THANK YOU