LEAD SCORE CASE STUDY

BY:
KAKSHI VILAS DONGRE

PROBLEM STATEMENT

X Education sells online courses to industry professionals. 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%.

BUSINESS GOAL

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score 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%

STEPS:

1. DATA CLEANING AND IMPUTING MISSING VALUES

2. EXPLORATORY DATA ANALYSIS : UNIVARIATE , BIVARIATE and MULTIVARIATE ANALYSIS

3. FEATURE SCALING AND DUMMY VARIABLE CREATION

4. LOGISTIC REGRESSION MODEL BUILDING

5. MODEL EVALUATION : SPECIFICITY , SENSITIVITY, PRECISION and RECALL

6. CONCLUSION AND RECOMMENDATION

DATA CONVERSION

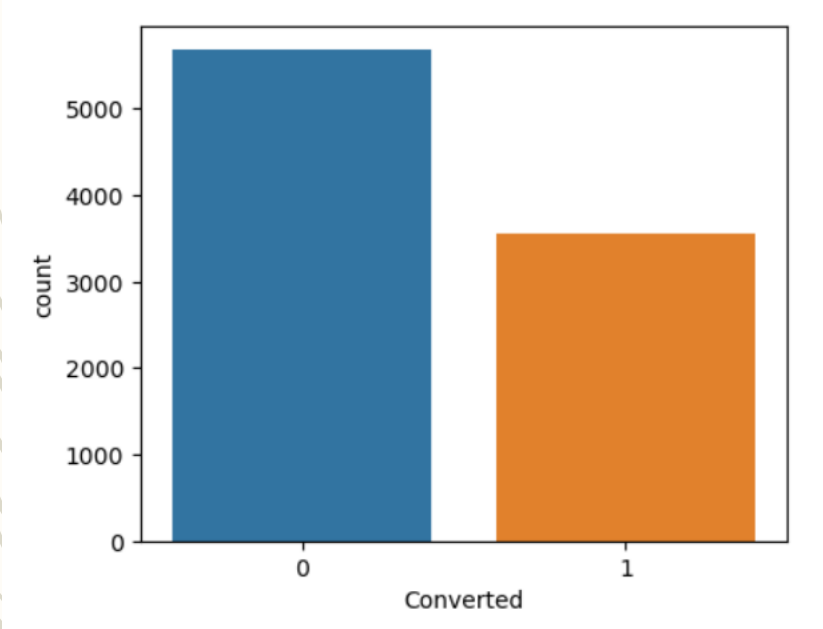
1. CONVERTING THE VARIABLE WITH VALUES YES/NO to 1/0s

2. CONVERTING THE ‘SELECT’ VALUES WITH NaNs

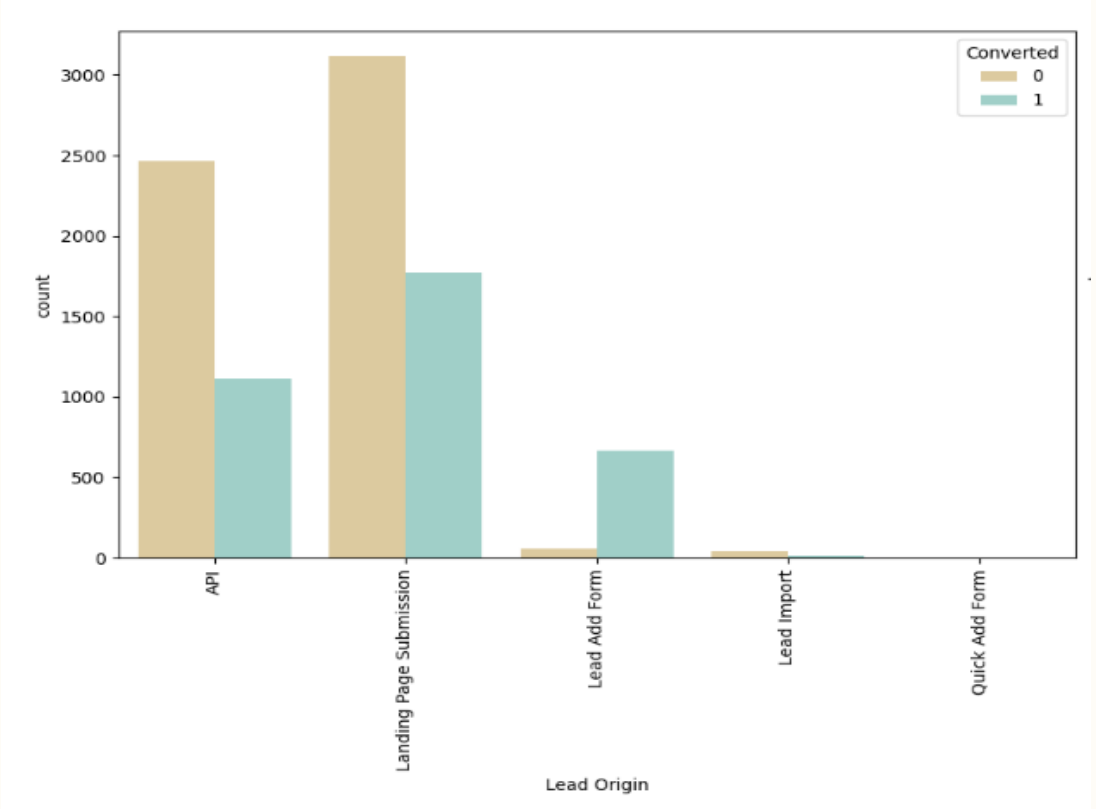
3. DROPIING THE COLUMNS HAVING >40% OF NULL VALUES

4. DROPPING UNNECESSARY COLUMNS

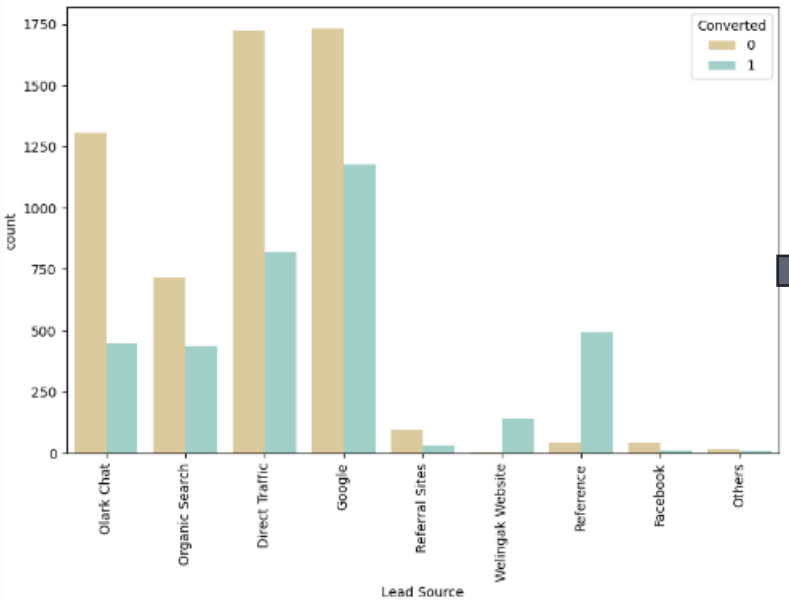
EDA



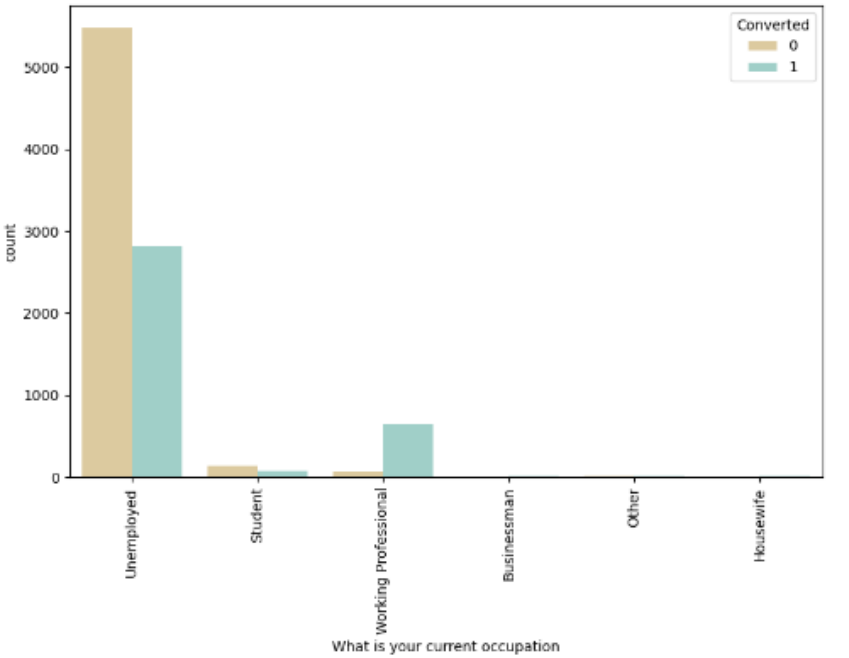
We have around 38%of Conversion Rate



1. API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable.
2. The count of leads from the Lead Add Form is pretty low but the conversion rate is very high

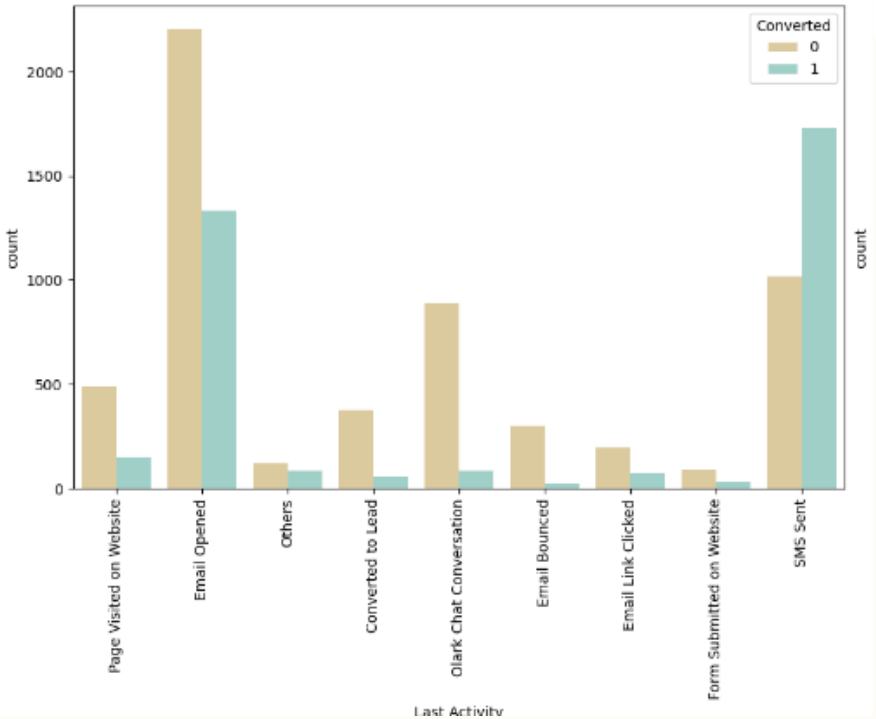


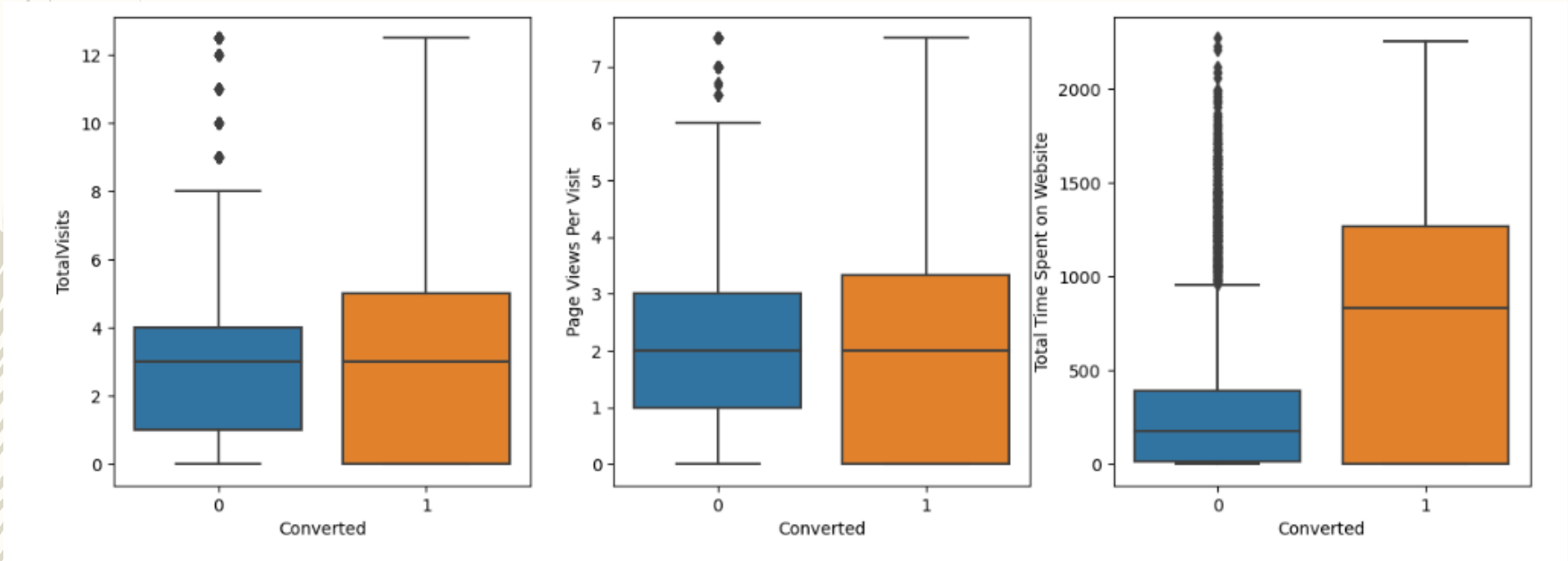
1. The count of leads from the Google and Direct Traffic is maximum.
2. The conversion rate of the leads from Reference and Welingak Website is maximum



1. Number of Unemployed leads are more than any other category.
2. Working professional has good conversion rate.

The conversion rate of SMS sent as last activity is maximum





The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information.
Users spending more time on the website are more likely to get converted

1. SPLITTING THE DATA INTO TEST AND TRAINING SETS.

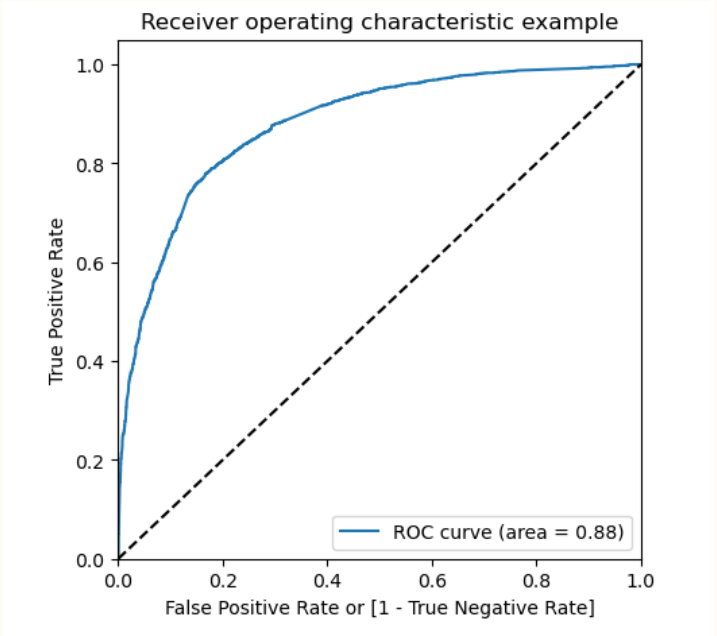
2. WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO AS 70:30.

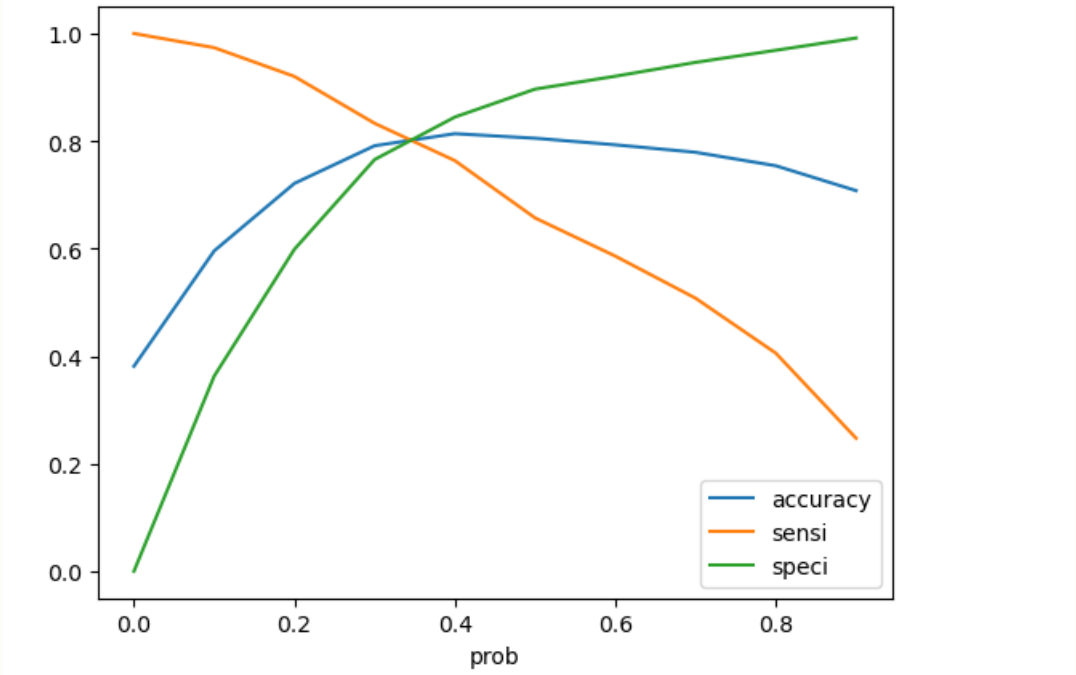
3. USING RFE TO CHOOSE TOP 15 VARIABLES.

4. PREDICTIONS ON TEST DATASET.

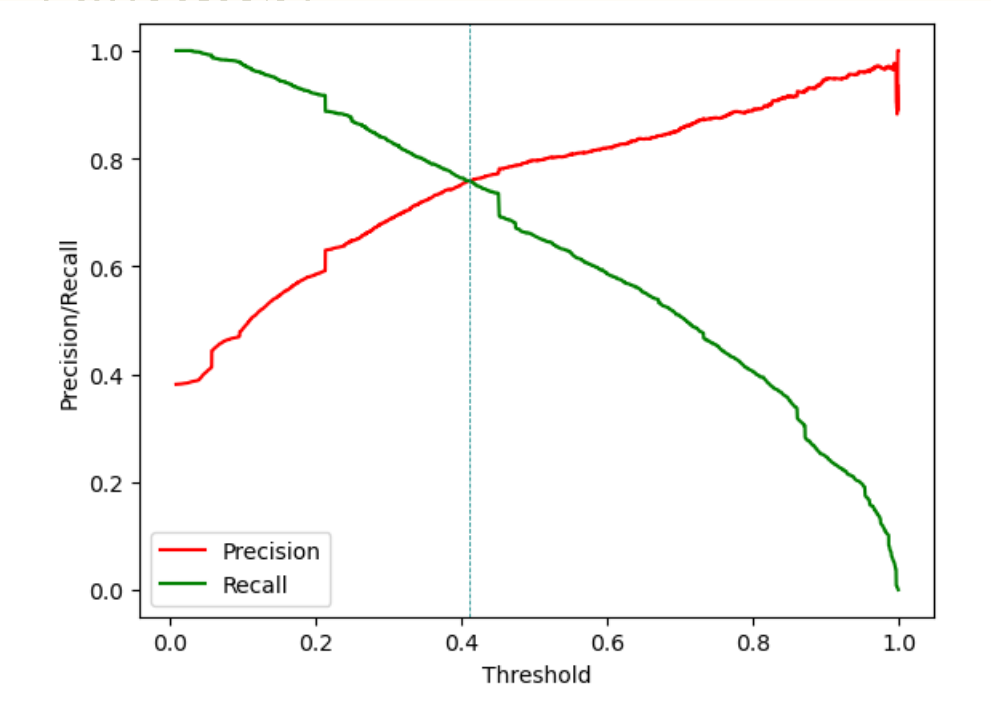
5. BUILD MODEL BY REMOVING THE VARIABLES WHOSE p-VALUE > 0.05 AND VIF > 5.

6. OVERALL ACCURACY IS 88%





PRECISION AND RECALL ON TRAIN DATASET



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

CONCLUSION

1. The logistic regression model is used to predict the probability of conversion of a customer.

2. Accuracy, Sensitivity and Specificity values of test set are which are approximately closer to the respective values calculated using trained set.

3. The top 3 variables that contribute for lead getting converted in the model are:
   - Lead Source_Welingak Website
   - Lead Source_Reference
   - Current_occupation_Working Professional

4. Hence overall this model seems to be good.