LEAD SCORING: ASSIGNMENT ON LOGISTIC REGRESSION

What is lead scoring?
Lead scoring is the process of assigning a “score” to each one of the contacts that reflects their conversion potential and level of interest in the stated business

Team members:

M D Shoaib

Mahesh B S

K .Thara Nayak

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. 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.

GOAL OF THIS ASSIGNMENT:

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 the 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.

STEPS FOLLOWED:

1. Inspecting and cleaning of the data set

- Checking for null values and dropping the values greater than 3000
- Dropping columns which do not require in the analysis- City, Country ,Prospect ID and lead Number, "What matters most to you in choosing a course"
- Analysing the values under significant columns like "Specialization", "How did you hear about X education" and dropping columns with higher "SELECT" as value
- We have dropped columns with only one value, null values

2. Exploratory Data Analysis:

- Univariate analysis- with countplots
- Bivariate analysis- Essentially using pair plots, box plots

3. With data prepared for modelling, we created dummy variables, split the data into train and test, scaled the variable using min-max

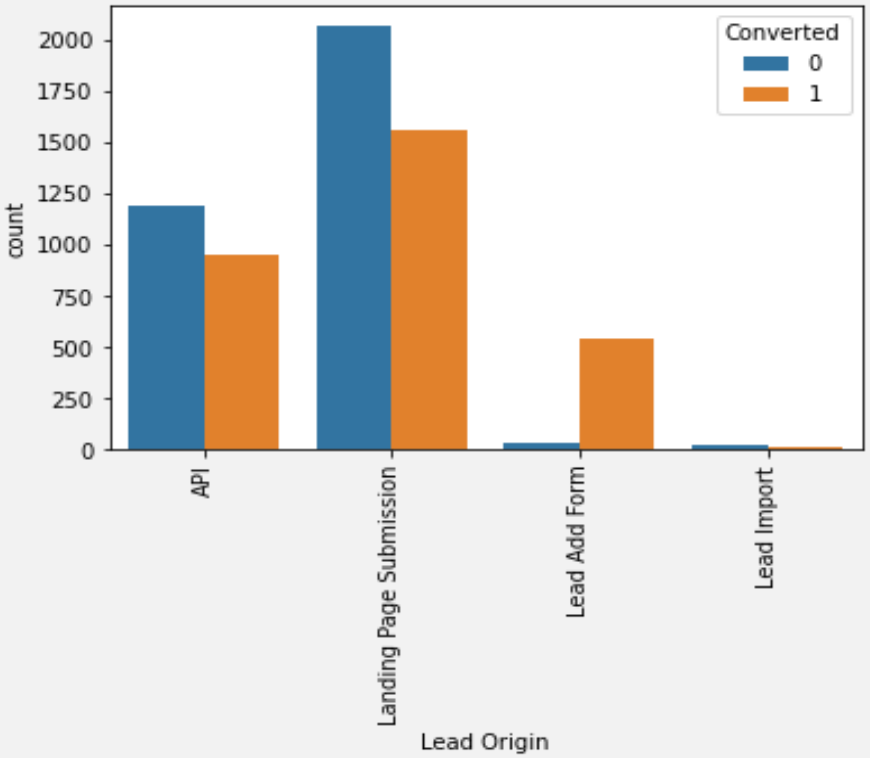
4. Model building using RFE

5. Checking of VIP, P-values until we achieved the accepted values

6. Plotted ROC curve, checked for necessary metrices- confusion matrix and the related metrics

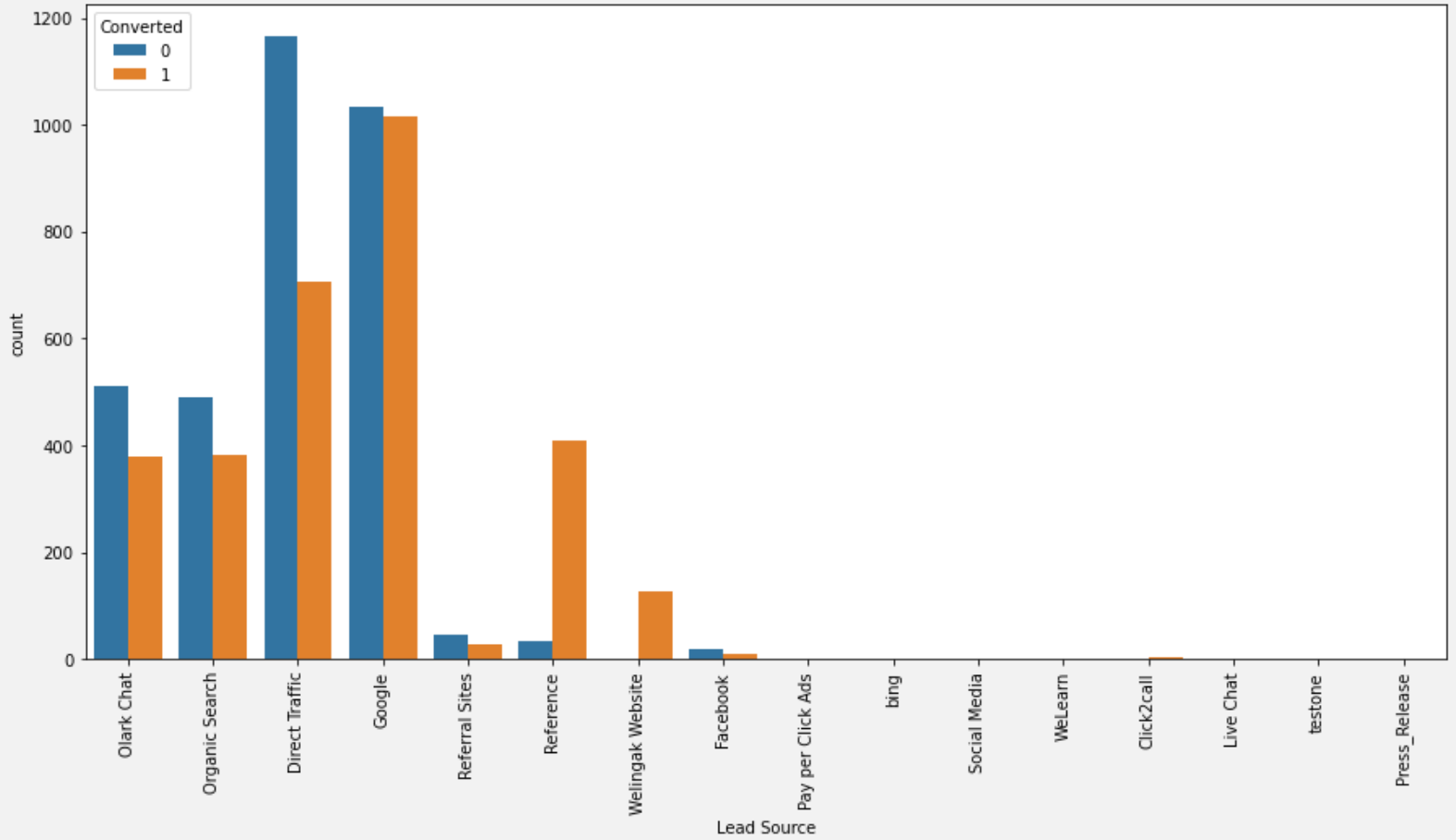
7. Made predictions on the test data

EDA



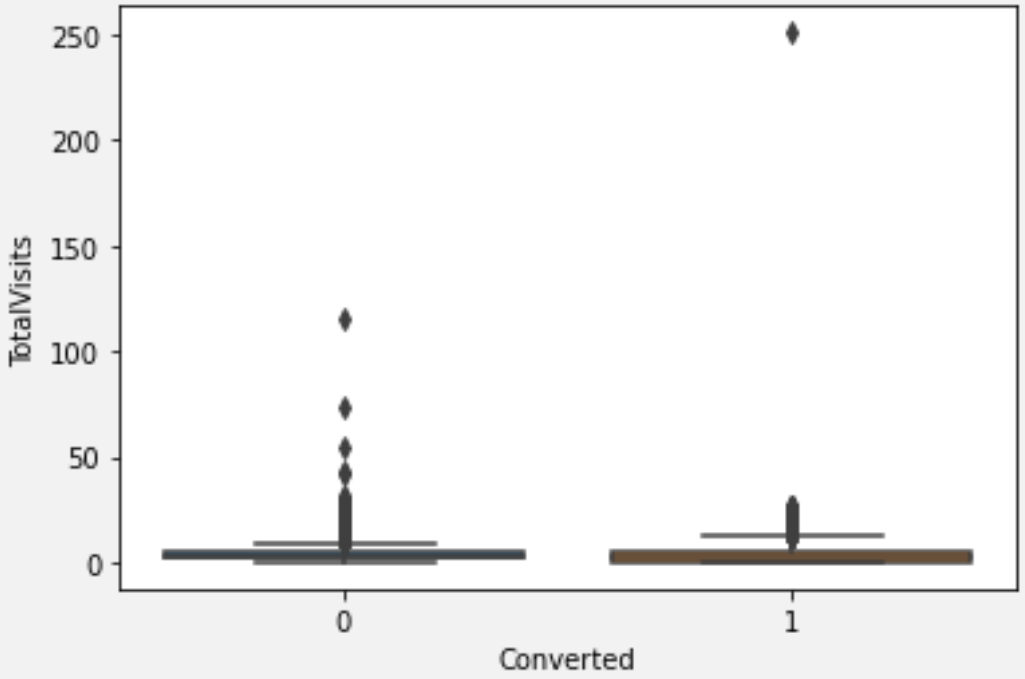
Lead origin:

API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable
Lead Add Form has more than 90% conversion rate but count of lead are not very high
Lead Import are very less in count.



Lead Source:

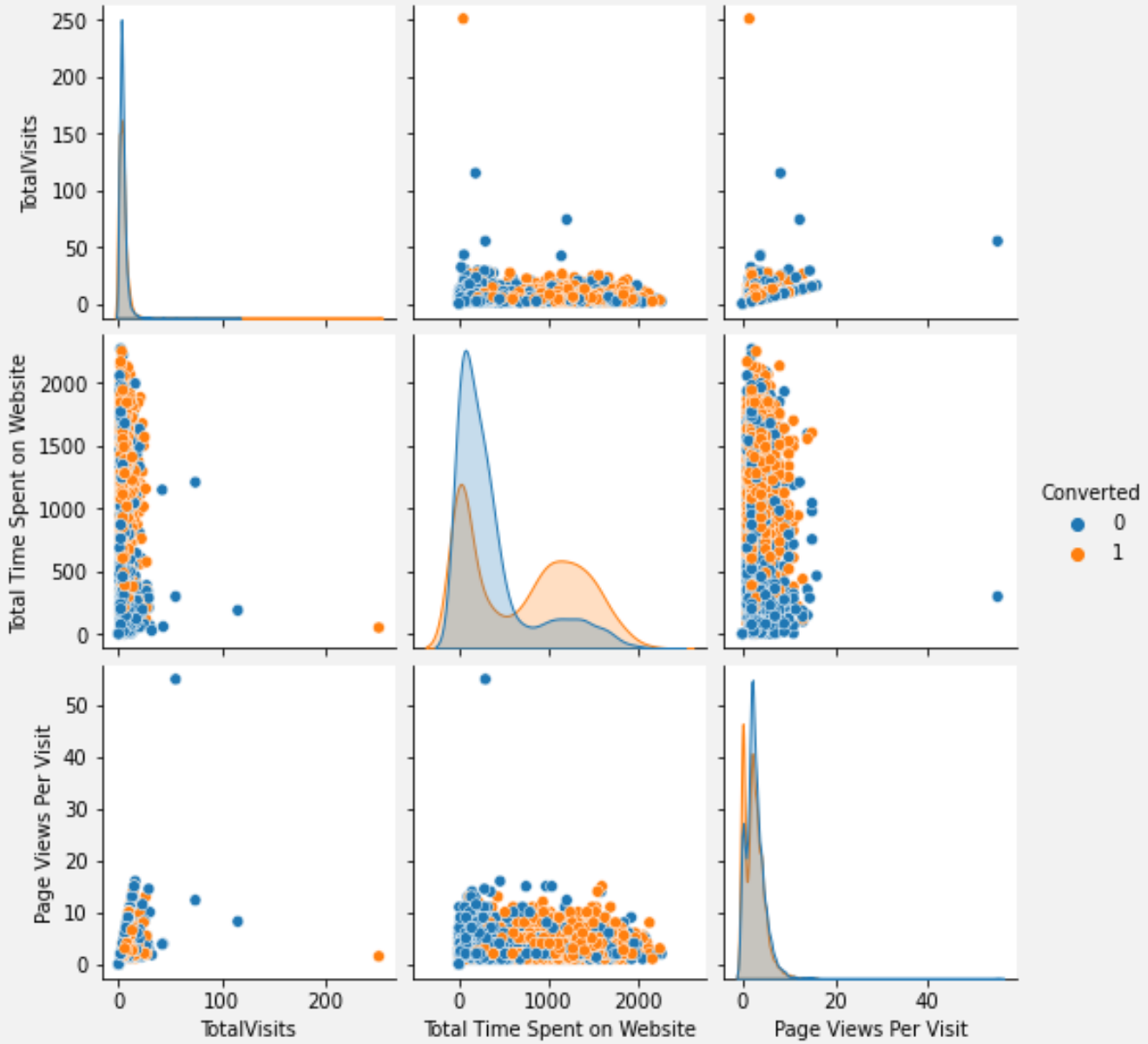
•Conversion Rate of reference leads and leads through Google and direct traffic is good
•. References and Organic search also show some conversion rate



Box Plot:

Total Visits on the converted data:

We see that the visit to website has one outlier of the value being 256. Median value however is almost hovering around the same value (somewhere around 3)



MODEL BUILDING:

Dummy variable creation

```
Index(['Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity',
       'Specialization', 'What is your current occupation',
       'A free copy of Mastering The Interview', 'Last Notable Activity'],
      dtype='object')
```

Scaling numeric variables

```
from sklearn.preprocessing import MinMaxScaler

Scaling numeric features

scaler = MinMaxScaler()
```

RFE

```
from sklearn.feature_selection import RFE
rfe = RFE(estimator= logreg,n_features_to_select=15)
rfe = rfe.fit(X_train, y_train)
```

Logistic Regression model fitting

```
X_train_sm = sm.add_constant(X_train)
logm2 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Final features after multiple iterations

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

MODEL EVALUATION:

Predicted probabilities on the train set

Created confusion matrix and derived the following metrices from it

- **Accuracy : 78.86%**
- **Sensitivity: 73.94%**
- **Specificity: 83.43%**
- ROC Curve:

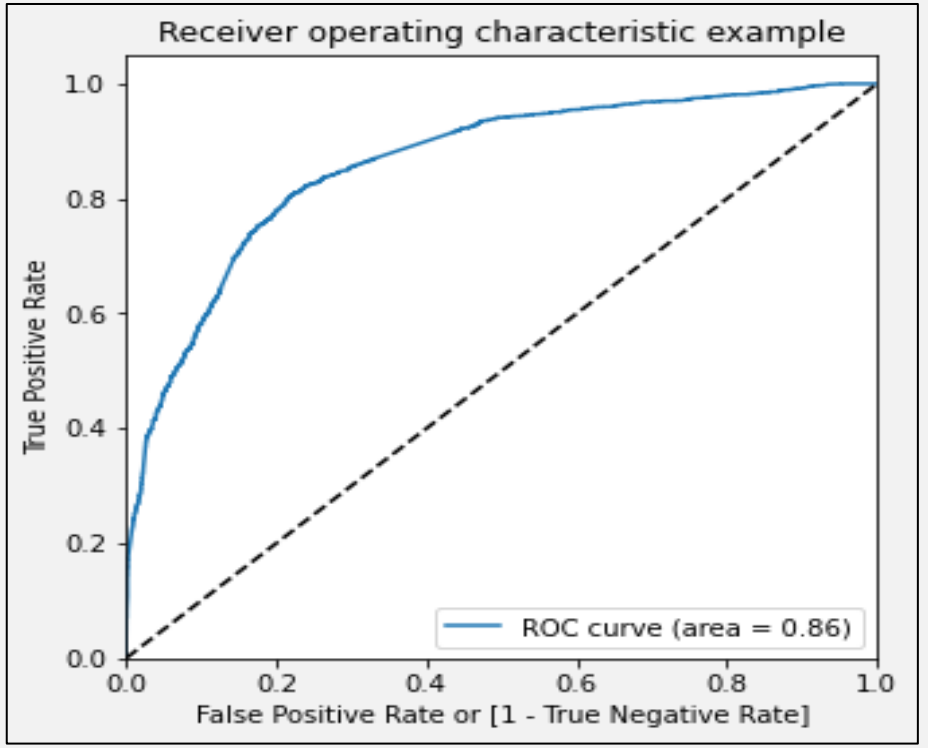
The more this curve is towards the upper-left corner, the more is the area under the curve (AUC) and the better is your model

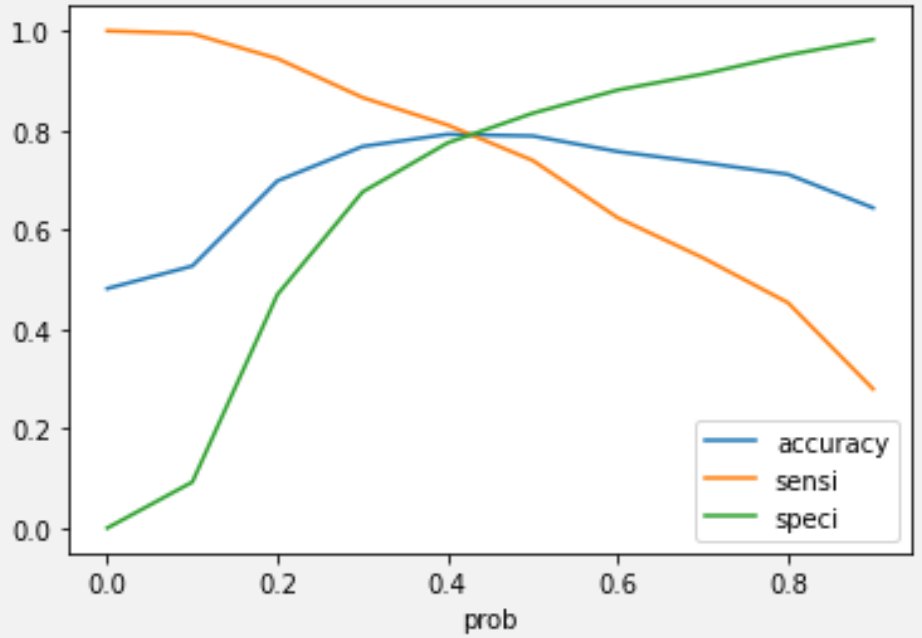
AUC for the curve which we plotted is **0.86**

We checked the sensitivity, specificity and accuracy trade-off to find the optimal cut-off probability point. And in our model it is somewhere around **0.42**

We recalculated the above metrices and they are as follows:

- **Sensitivity: 79.33%**
- **Specificity: 78.84%**





MAKING PREDICTIONS ON TEST DATA

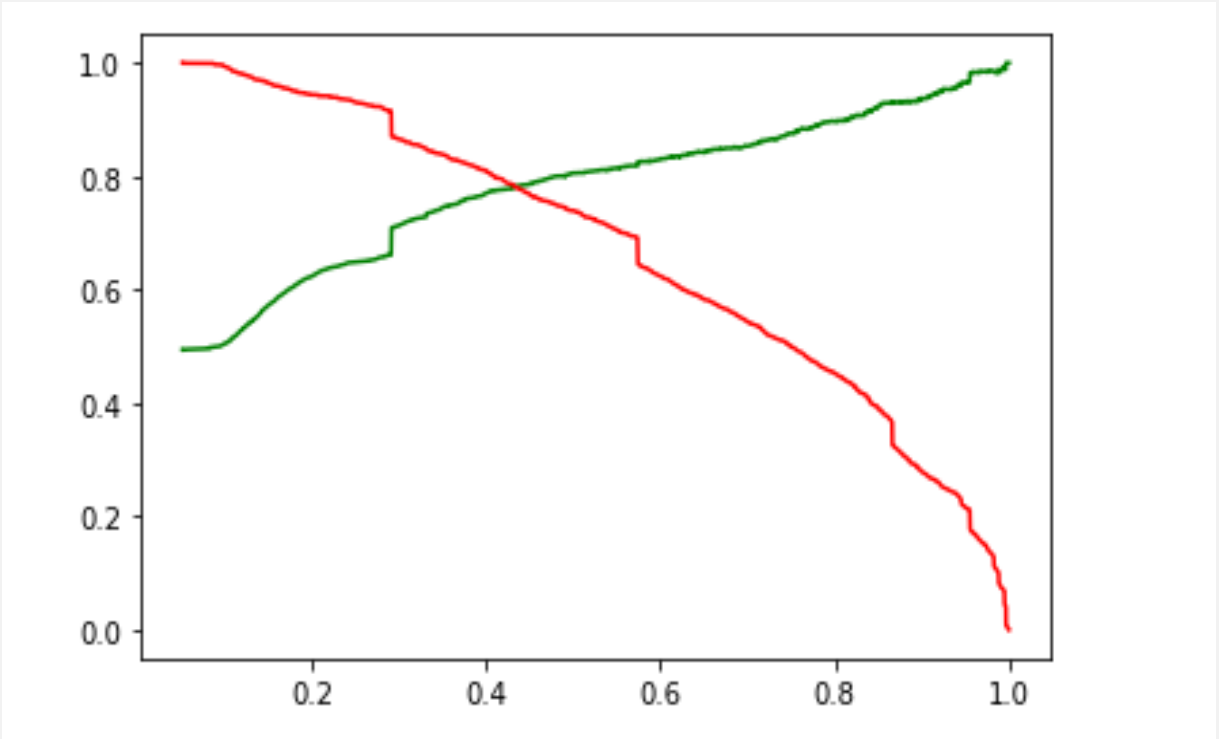
- Predictions were made on test data using the probability of 0.42
- Created confusion matrix and derived the following metrices from it

Sensitivity: 77.94%
Specificity: 78.91%

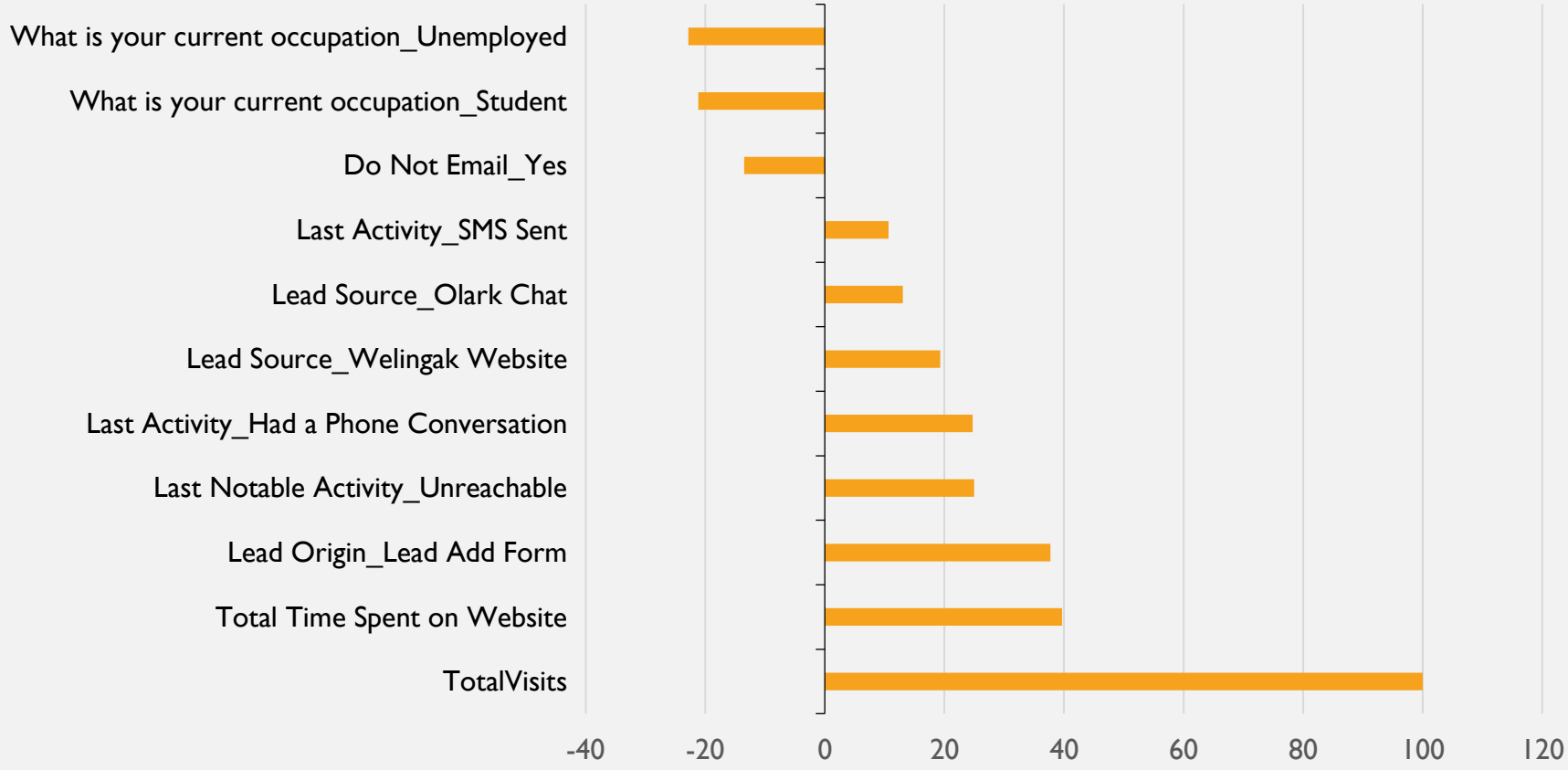
- Precision and Recall were calculated

Precision: 78.40%
Recall: 77.71%

- From the plotted curve, the cut off probability is around 0.44
- We now made predictions based on the probability=0.44 and our precision and recall score came up to 78.25% and 76.74% respectively



DETERMINING FEATURE IMPORTANCE



11 features are finalized for our model.

- Based on the coefficient values we have arrived at the feature importance
- Features with high positive coeff values are the ones that contribute most towards the probability of a lead getting converted.
- •Similarly, features with high negative coeff values contribute the least.

SUMMARY AND INSIGHTS

- Overall accuracy on Test set: 0.786
- Sensitivity of our logistic regression model: 0.7794
- Specificity of our logistic regression model: 0.7891
- Some of the features with positive coefficient values are as follows: Total Visits, Total time spent on the website, Lead origin(Add form), Lead source(Olark chat and Welingak website), last activity (Phone conversation and SMS sent)

- Features with negative coefficients are: Do not email(yes), Student's current occupation, Current employment status (Unemployed)

- Top 3 variables in model, that contribute towards lead conversion are:
  -Total Time Spent on Website.
  -Total Visits.
  -Lead Source with elements Google/organic search

- The recall here is 77.71%. Basically the model is 77.771% good in predicting positive class which is hot leads.

X Education company can consider the following to reach a good conversion rate of their leads to customer

- Focus more on current occupation anything but "Student" and "Unemployed". Conversion here is tacky. They have low potential for getting converted as customers for the company
- Approaching leads more through SMS and Olark chat can gain some significant conversion. These chat or SMS programs can be pre-designed that can help save the counsellor's time /effort.
- Leads from ad forms could also be focussed since they could also show a good conversion rate
- Increase user engagement on their website since this helps in higher conversion

THANK YOU