Lead Scoring Case Study

Presented By

1)Mohit Daga

2) Nilesh Borade

Problem Statement:

X Education, sell online courses

X Education Websites

People land on the website, If they fill up a form or drop Phone Number

they are classified to be a lead.

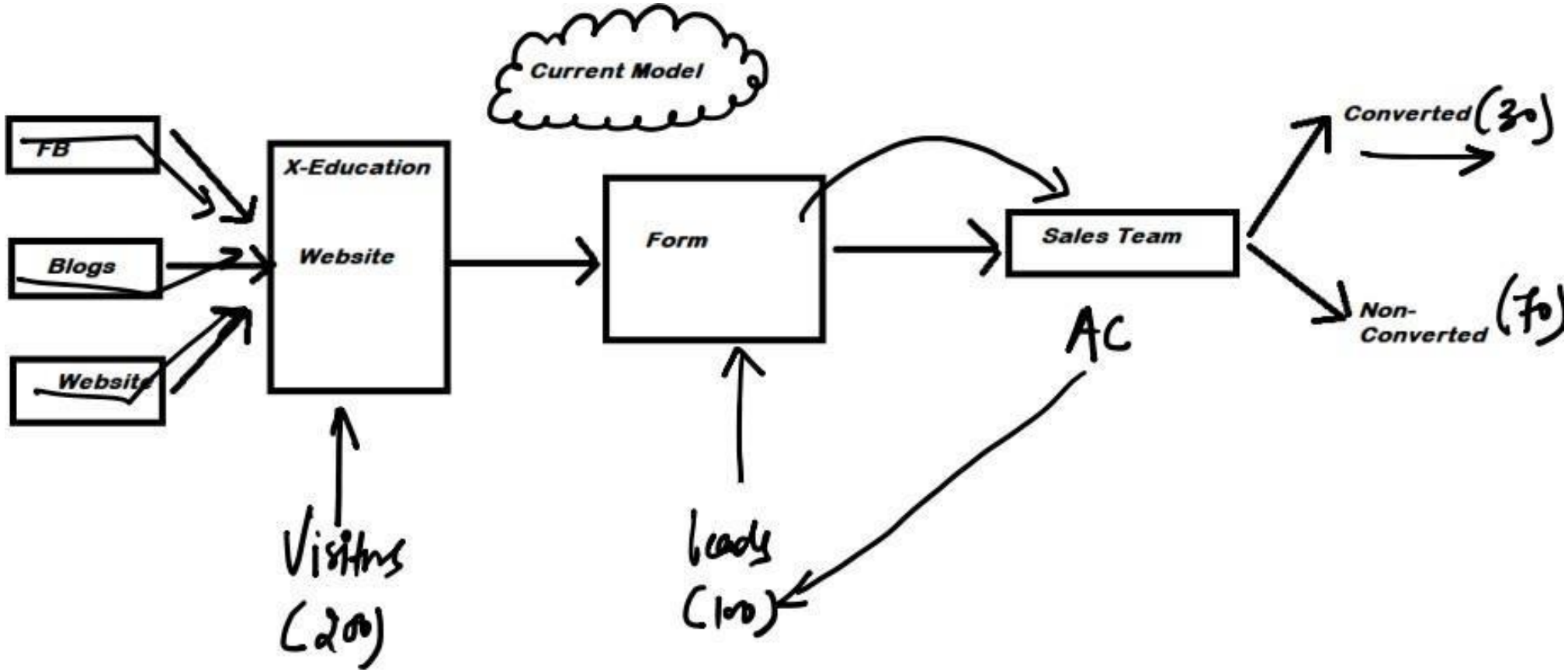
Oe sales team start making calls, writing emails, etc. to every lead

Some of the leads get converted into customers

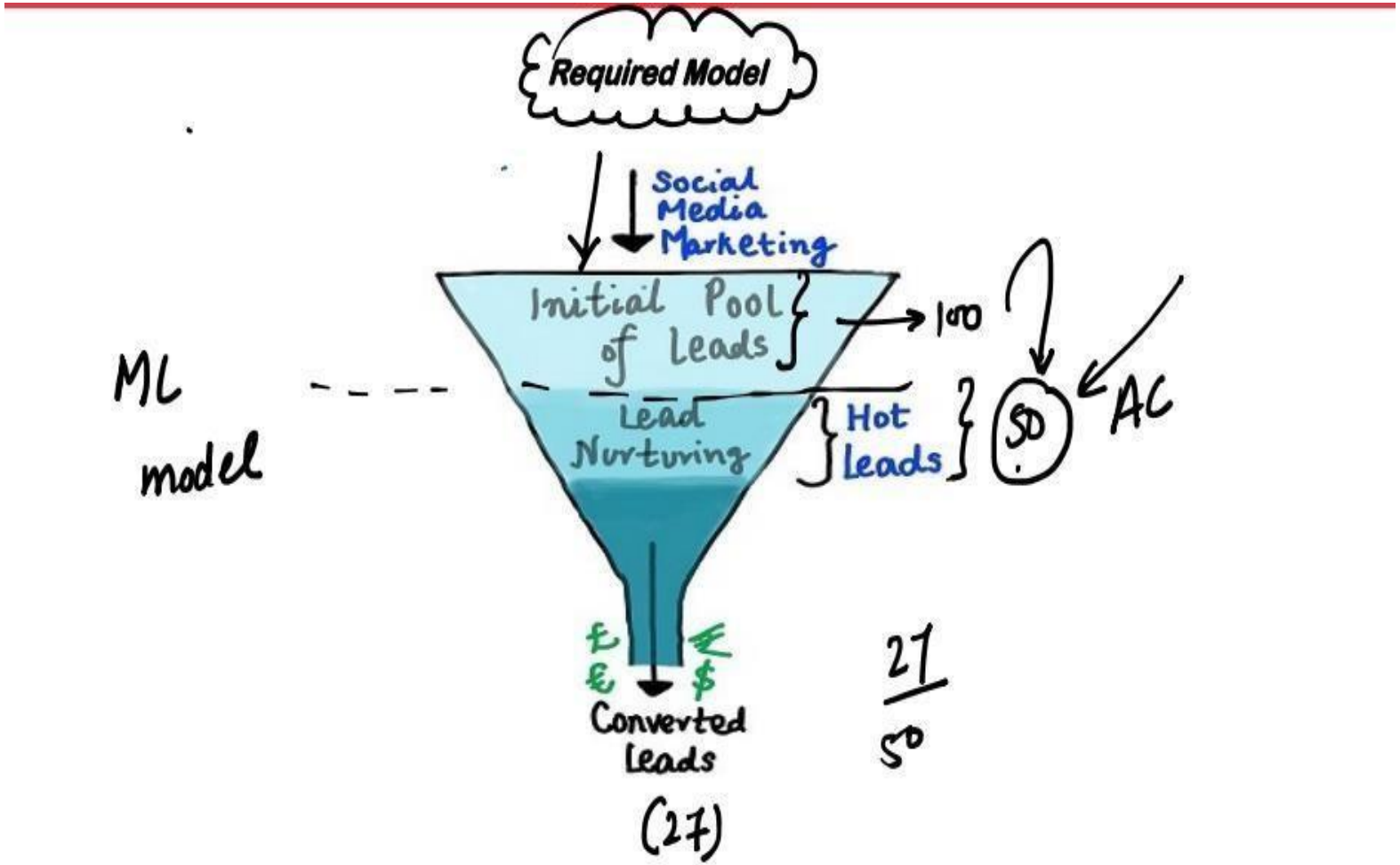
But here rate of conversion is very less, almost only 30%

The company wishes to identify the most potential leads, also known as ‘Hot Leads’. Focus on them only So most of conversion happens (Expecting 80%)

This is how current model looks



What they are expecting to build



Our Analysis Approach

- As we aim to predict weather lead is potential lead or not so it's becomes Classification Problem and thus we are using logistic regression model here
- So we can proceed with analysis in following steps

✓ Read & Understand Data

✓ Data Cleaning

Data Preparation

✓ Modelling

✓ Evaluation of Model

Step 1 : Read & Understand Data

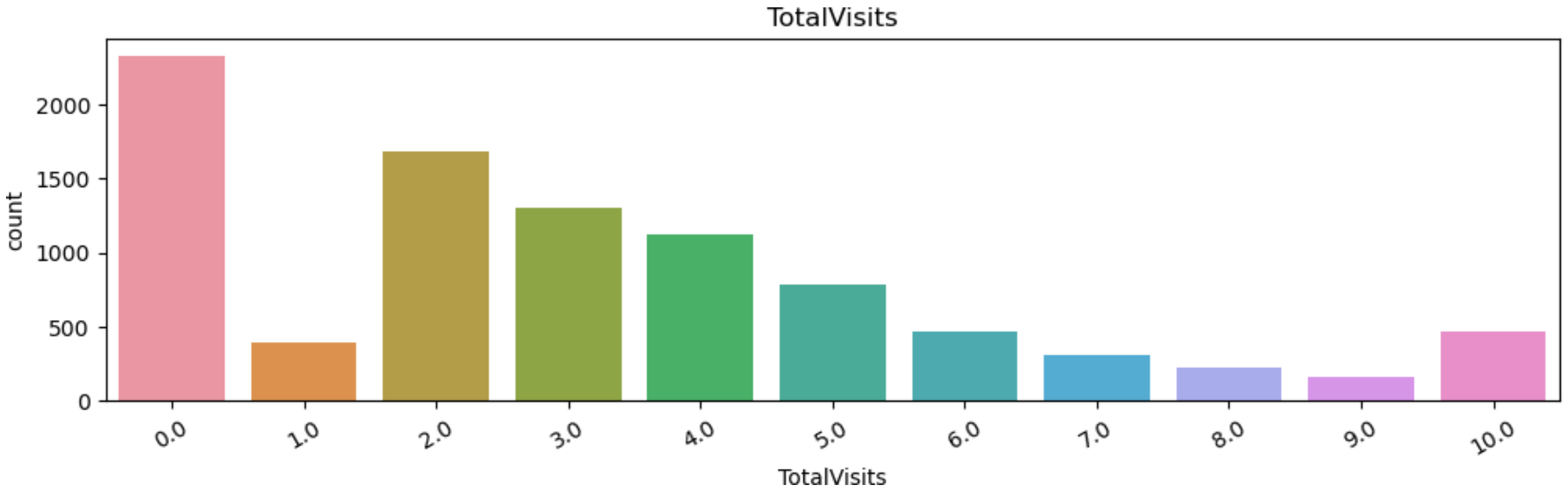
With the very first step, we started understanding 'Problem Statement'

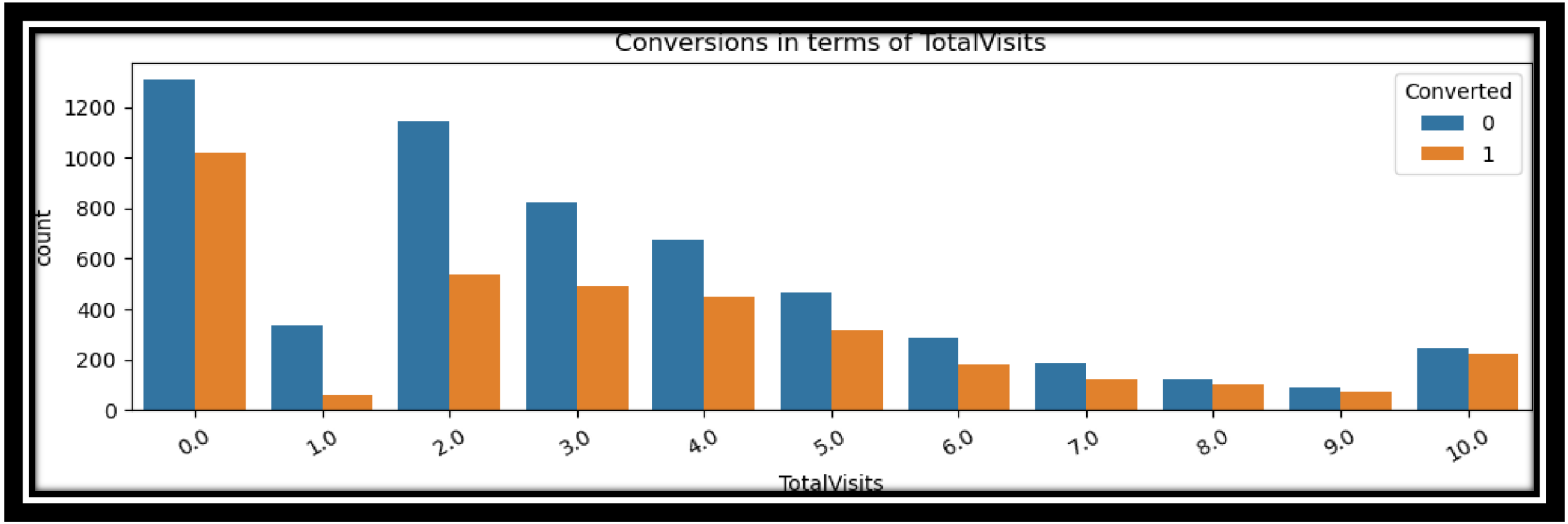
Then for to read & understand data imported all the required library in jupyter note book & read file, data dictionary help us to understand meaning of data in data file

Step 2: Data Cleaning

- Lot of variables has 'Select' values which are considered to be null values so it replaced with Nan.
- Columns have large number of null values i.e. greater then 40% so we dropped it.
- Since 'Project ID' and 'lead Number' are of no use in regression model and also all have unique values we dropped them.
- Still there was columns with missing value percentage between 25-40% so checked every column one by one & do necessary cleaning as you can see in jupyter file
- So as part of cleaning data we drop some unnecessary column, binning, handle outlier, imputing & so on

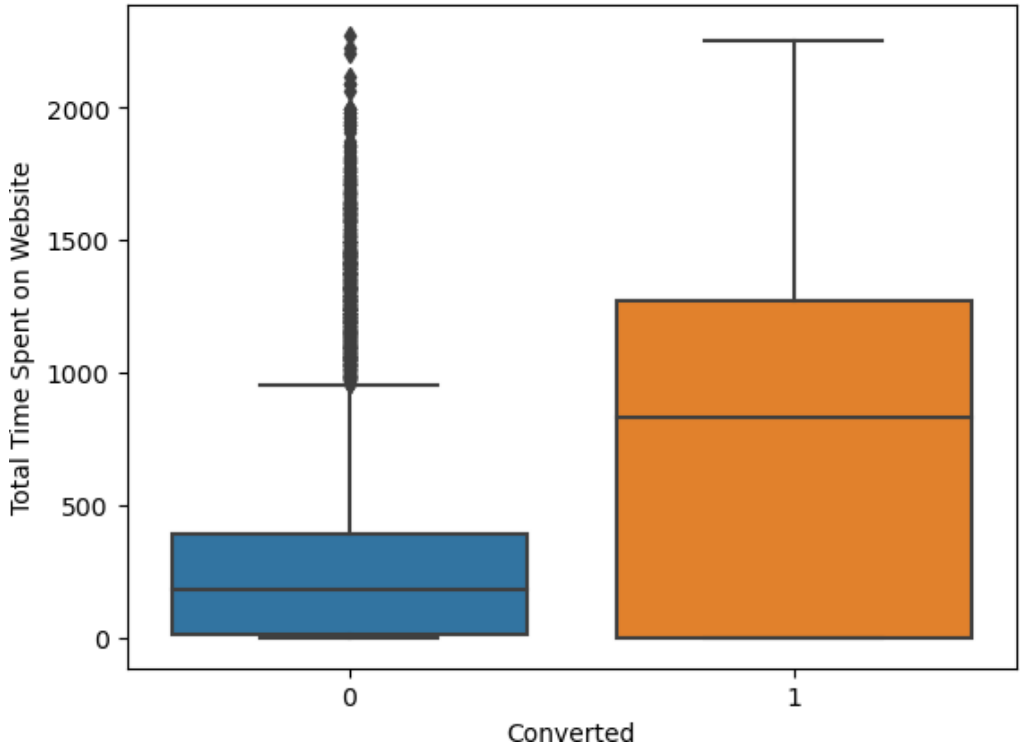
- **Total Visits**





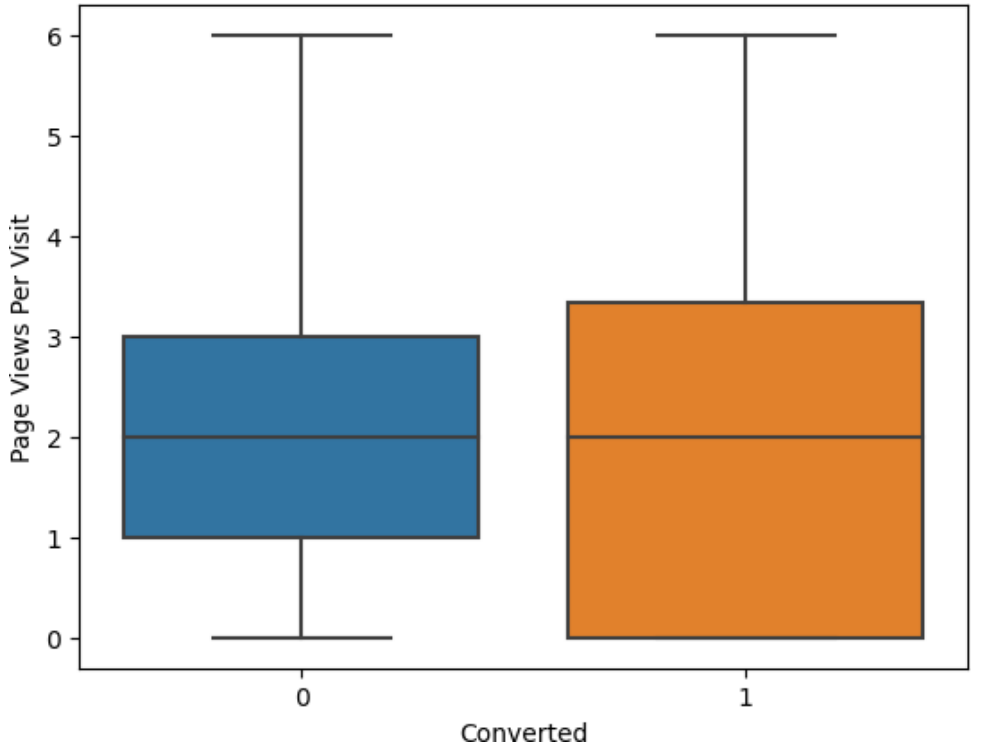
- Most of the leads have not visited, but have highest rate of conversion, as well when lead visited 8-10 times most of time it get converted

• Total Time Spent on Website



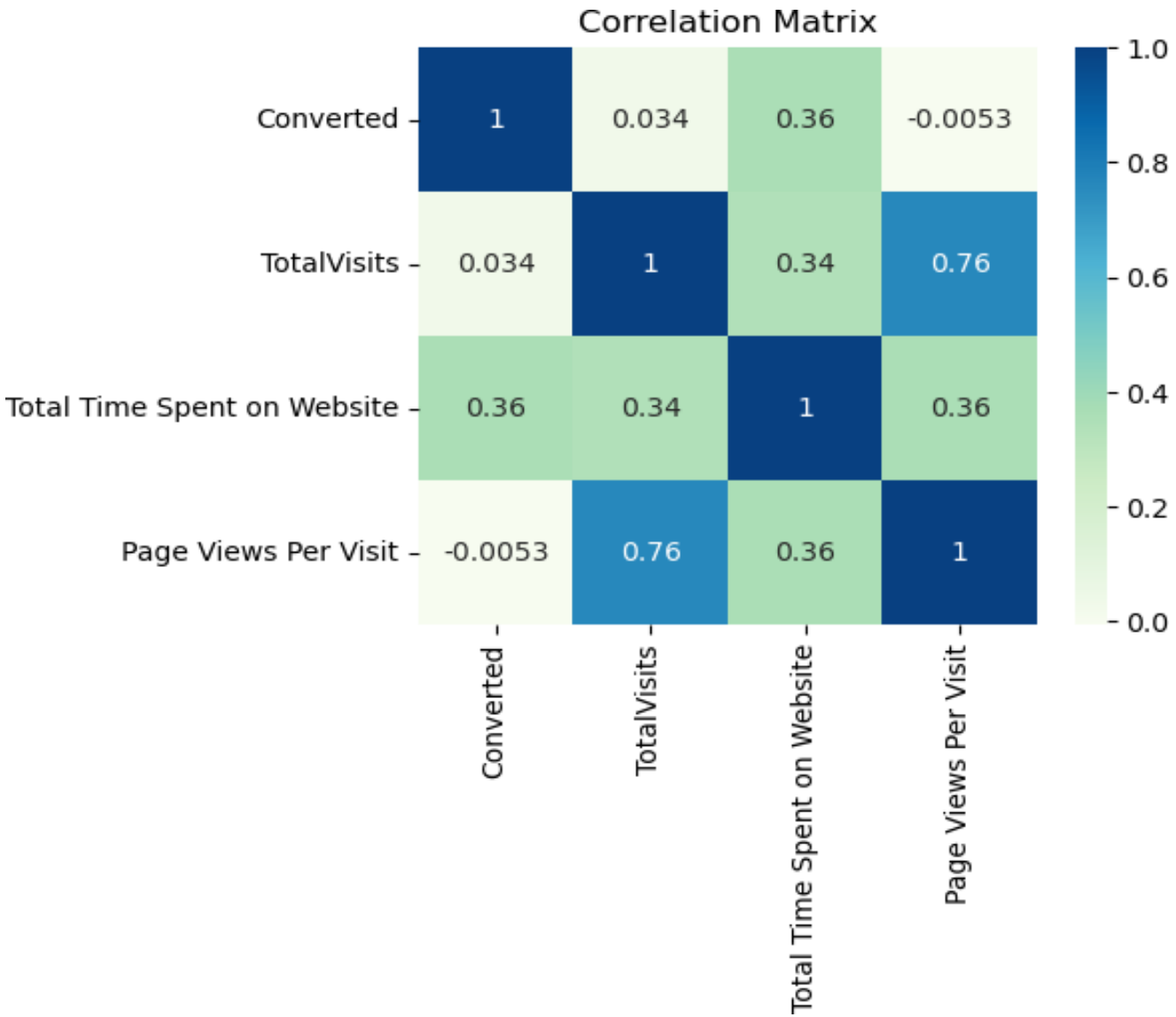
- This shows that more the time spent on website more are the leads that got converted.
- People who have spent less time have not converted.

• Page Views Per Visit



- Median for both the opted and not opted is same.
- people who have visited 1-3 on an avergare have 50-50 percent chances of getting converted.

Checking for numerical variables for collinearity



- Since we have high collinearity between Page views per visit and Total visitors we can drop any one of them.

Step 3: Data Preparation

- Here we convert all data into numeric form (like Yes/No into 1/0)
- Create dummies for all categorical variables
- Performed train-test split &
- Performed scaling using Sklearn Library

Step 4: Model Building

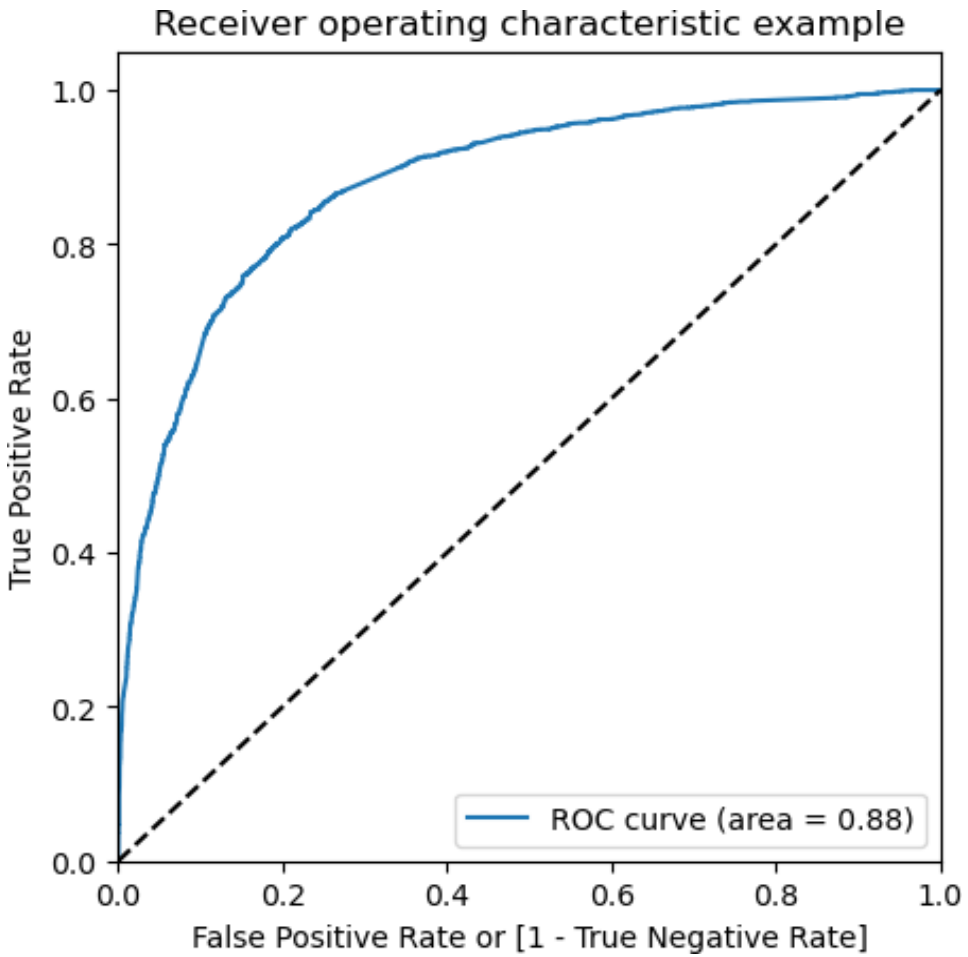
- Here we aim to build logistic regression model
- So we started with feature selection using RFE
- With the help of P value & VIF we reached to most suitable model

Step 5: Model Evaluation

- Evaluating model with following parameter

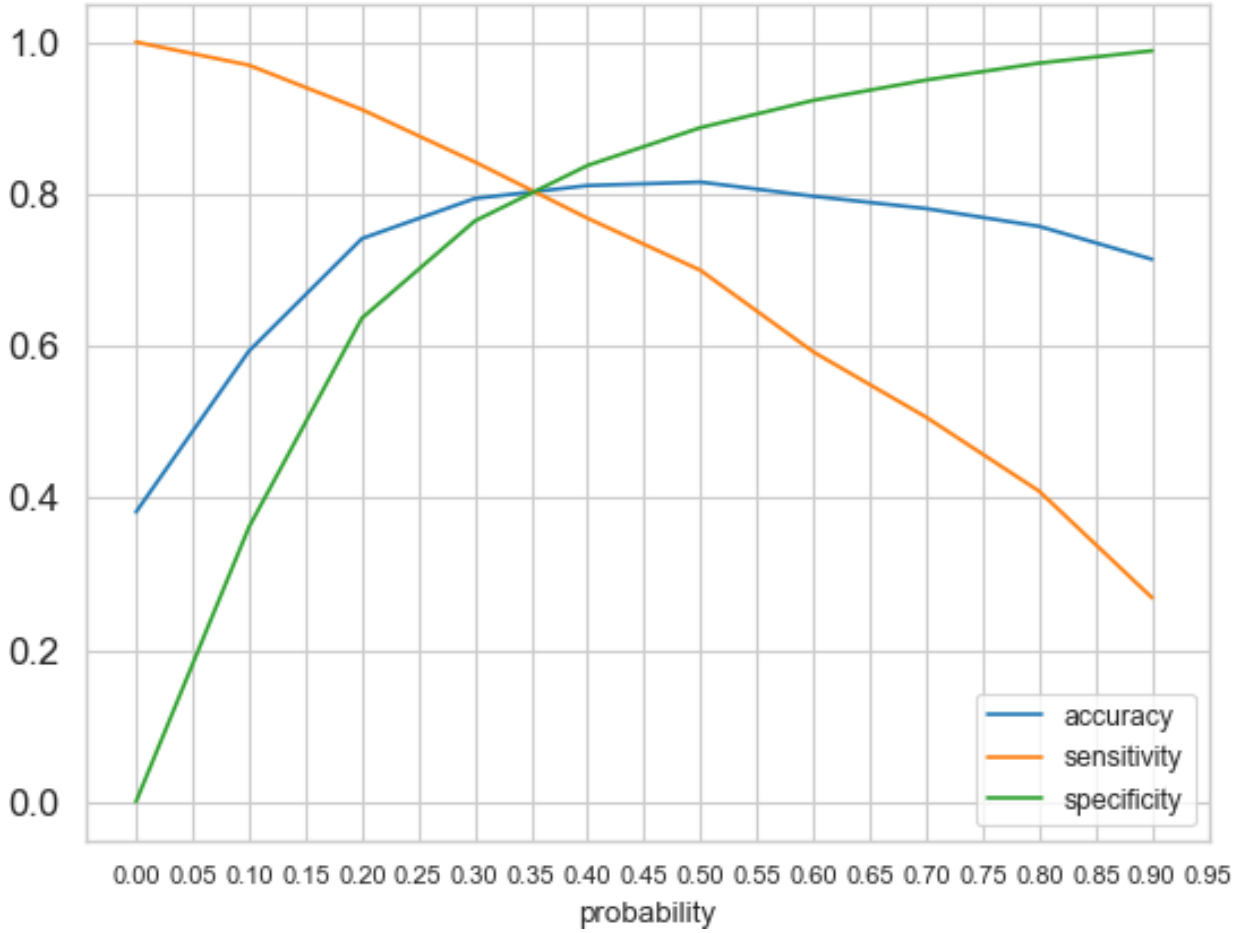
Sr.No / Parameter	Parameter	Value
1	Accuracy	0.8157
2	Sensitivity	0.6995
3	Specificity	0.8873
4	False positive rate	0.1126
5	Positive predictive value	0.7927
6	Negative predictive value	0.8273

Plotting the ROC curve



- The ROC Curve should be a value close to 1. We are getting a value of 0.88 indicating a good predictive model.
- AUC stands for Area under the ROC curve shows that 88% of the predictions are correct. hence more the AUC better the model.

Finding Optimal Cutoff Point



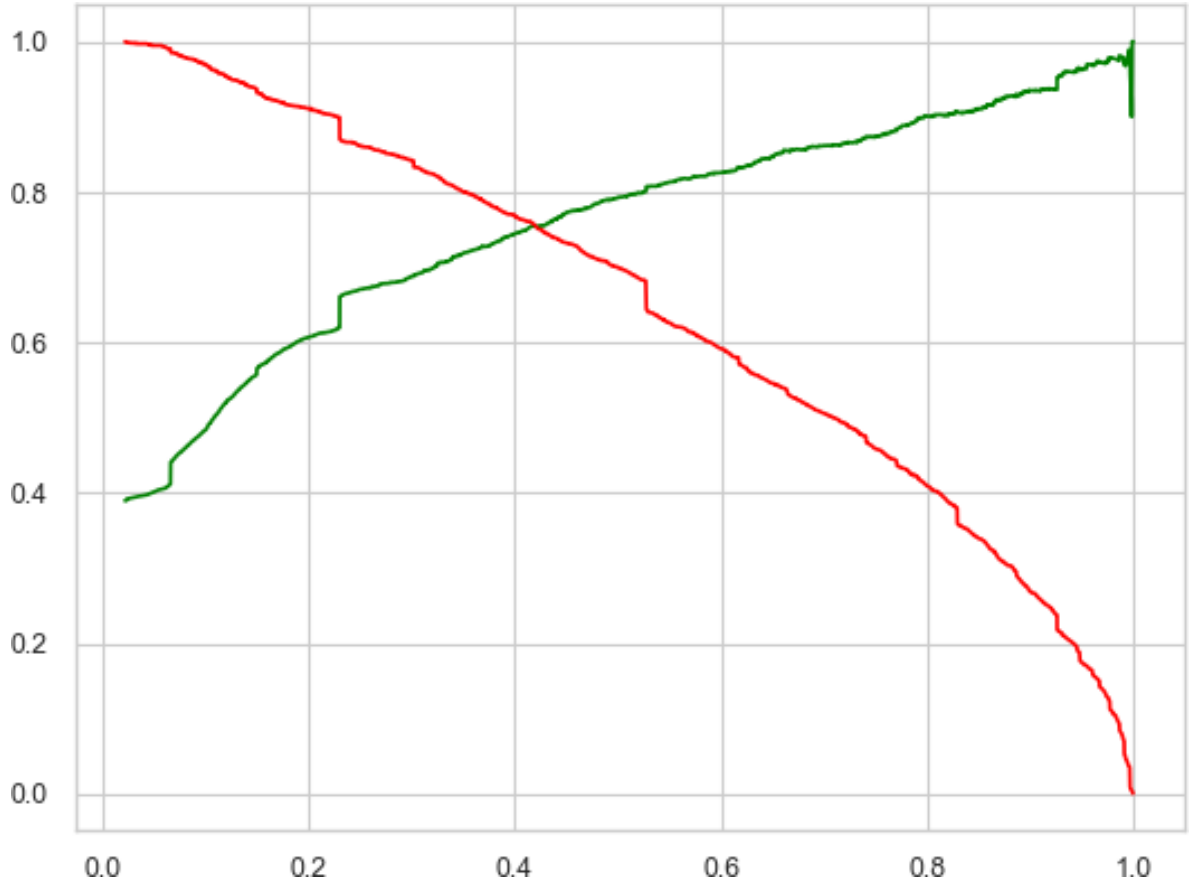
- From the curve, 0.35 is the optimum point to take it as a cutoff probability.

- It means that at this point Accuracy, Sensitivity and Specificity are all same.

Precision and Recall

- **Precision**
  - Precision measures how good our model is when the prediction is positive.
  - We need high precision in places such as recommendation engines, spam mail detection, etc. Where you don't care about false negatives but focus more on true positives and false positives. It is ok if spam comes into the inbox folder but a really important mail shouldn't go into the spam folder.
  - TP / TP + FP
  - **We got Precision Value `0.69`**
- **Recall**
  - measures how good our model is at correctly predicting positive classes.
  - Models need high recall when you need output-sensitive predictions. For example, predicting cancer or predicting terrorists needs a high recall, in other words, you need to cover false negatives as well. It is ok if a non-cancer tumor is flagged as cancerous but a cancerous tumor should not be labeled non-cancerous.
  - TP / TP + FN
  - **We got Recall Value `0.84`**

Precision and Recall tradeoff



- This graph shows that Precision and Recall are inversely related as one increases other decreases.

Final Observation

We make prediction on test data set & got following results compare to train data set

- **Train Data**
- Accuracy - 80.00%
- Sensitivity - 84.22%
- Specificity - 76.43%
- Precision - 69.00%
- Recall - 84.22%

- **Test Data**
- Accuracy - 81.27%
- Sensitivity - 84.22%
- Specificity - 76.43%
- Precision - 77.00%
- Recall - 75.06%

Conclusion

- We got around 1% difference on train and test data's performance metrics.This implies that our final model didn't overfit training data and is performing well.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.
- Depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.

Recommendations:

1. **Total Time Spent on Website**
   - From The Boxplot we can see that number of constumers who have converted spent more time on the website.so try to make them engage in the website and increase their chances of getting converted.
   - Try to give them offers exclusive on websites so they gets attracted to it.

2. **Lead Source**
   - *Olark Chat:* We can see that olark chat has majority of lead conversions in terms of count, so focus on increasing the lead conversions to this categories.
   - *Welingak Website:* Welingak website has higher number of lead conversion rate but count is low.so you can improve and learn from here that might help you in improving other categories.

1. **Lead Origin**
   - *Lead Add Form:*
   - Lead add form has higher number of lead conversion rate so you can improve and learn from here that might help you in improving other categories.

6. Last Notable Activity

- SMS Sent:
  - SMS sent has good conversion rate and so there is chance of improvance more.
- Others:
  - Others has many categories such as 'Unreachable','Unsubscribed','Email Bounced','Had a Phone Conversation', 'Approached upfront','View in browser link Clicked','Email Received','Email Marked Spam','Form Submitted on Website', 'Resubscribed to emails'.
  - Since the lead count and converted count is less try on increasing the number as your
    - leads score depends on this


THANK YOU