Lead Scoring Assignment

By Arpana, Jayanand and Rajesh

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 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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

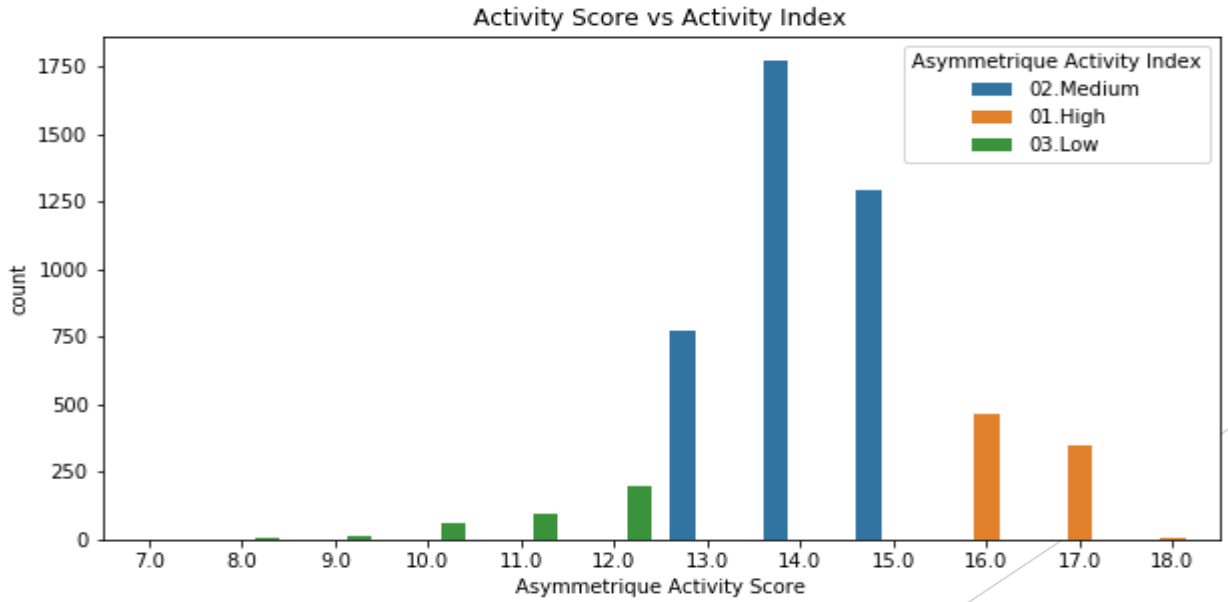
Fig. Lead Conversion Process

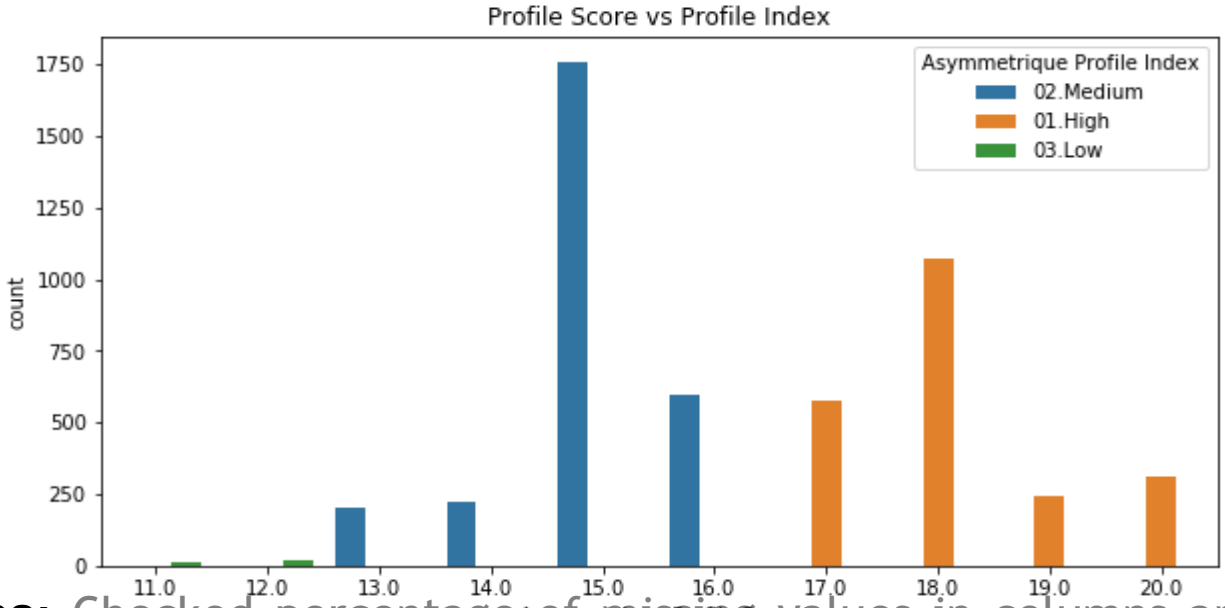
Objectives

- Help the company by building a model to select the most promising leads, i.e. the leads that are most likely to convert into paying customers to assign a lead score 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. To help the company in selecting the most potential leads, also known as ‘Hot Leads’ whose lead conversion rate is around 80%.
- 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., i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Company requires model that should be able to adjust if the company’s requirement changes in near future
- This helps the sales team to divert their focus on potential leads

Approach

- In application new dataset we have 37 columns and 9240 rows
- Reading and Understanding the dataset
- **Data cleaning:** We started our analysis with our cleaned dataset by dropping unnecessary data , converting all the binary variables to ‘0’ and ‘1’ and multiple categories into dummy variables.
- Next, we checked the outliers of the dataset. The visualization of those outliers we can see on the graph attached on the right side.
- understand the relationship between the score and index





Handling Missing values: Checked percentage of missing values in columns and dropped the columns which h missing values more than 40% .

Group the similar specialisations into a category to gain more meaningful insights

There are days columns which is having NAN values, so we are converting that to 'Others'.

Handling duplicates by dropping them.

Created dummy variables and dropping the first column

Model Building

Model 1

Dep. Variable:	Converted	**No. Observations:**	6194
Model:	GLM	**Df Residuals:**	6178
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1336.7
Date:	Sat, 25 Feb 2023	**Deviance:**	2673.3
Time:	13:46:46	**Pearson chi2:**	7.44e+03
No. Iterations:	20	**Covariance Type:**	nonrobust

	Features	VIF
14	Last Notable Activity_SMS Sent	6.06
4	Last Activity_SMS Sent	6.04
1	Lead Origin_Lead Add Form	1.83
12	Last Notable Activity_Modified	1.73
11	Tags_Will revert after reading the email	1.68
2	Lead Source_Olark Chat	1.65
10	Tags_Others	1.65
0	Total Time Spent on Website	1.47
3	Lead Source_Welingak Website	1.33
5	Tags_Closed by Horizzon	1.21
13	Last Notable Activity_Olark Chat Conversation	1.07
7	Tags_Lost	1.06
9	Tags_Not Eligible/Not Interested	1.04
8	Tags_No Phone Number	1.02
6	Tags_Lateral student	1.01

Model 2

Dep. Variable:	Converted	**No. Observations:**	6194
Model:	GLM	**Df Residuals:**	6181
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1524.1
Date:	Sat, 25 Feb 2023	**Deviance:**	3048.2
Time:	13:46:54	**Pearson chi2:**	6.61e+03
No. Iterations:	8	**Covariance Type:**	nonrobust

	Features	VIF
1	Lead Origin_Lead Add Form	1.810657
2	Lead Source_Olark Chat	1.621037
8	Tags_Others	1.540501
0	Total Time Spent on Website	1.465145
10	Last Notable Activity_Modified	1.454311
9	Tags_Will revert after reading the email	1.399377
3	Lead Source_Welingak Website	1.327094
4	Tags_Closed by Horizzon	1.211620
11	Last Notable Activity_Olark Chat Conversation	1.061252
5	Tags_Lost	1.058185
7	Tags_Not Eligible/Not Interested	1.036098
6	Tags_No Phone Number	1.005999

Model 3

Dep. Variable:	Converted	**No. Observations:**	6194
Model:	GLM	**Df Residuals:**	6182
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1527.9
Date:	Sat, 25 Feb 2023	**Deviance:**	3055.9
Time:	13:47:01	**Pearson chi2:**	6.69e+03
No. Iterations:	8	**Covariance Type:**	nonrobust

	Features	VIF
1	Lead Origin_Lead Add Form	1.810393
2	Lead Source_Olark Chat	1.620600
7	Tags_Others	1.529101
0	Total Time Spent on Website	1.465057
9	Last Notable Activity_Modified	1.406748
8	Tags_Will revert after reading the email	1.398212
3	Lead Source_Welingak Website	1.327067
4	Tags_Closed by Horizzon	1.209738
10	Last Notable Activity_Olark Chat Conversation	1.061018
5	Tags_Lost	1.056865
6	Tags_No Phone Number	1.005801

Getting predicted values on train dataset

Created new column 'predicted' with 1 if Prob > 0.5 else 0

We got below results

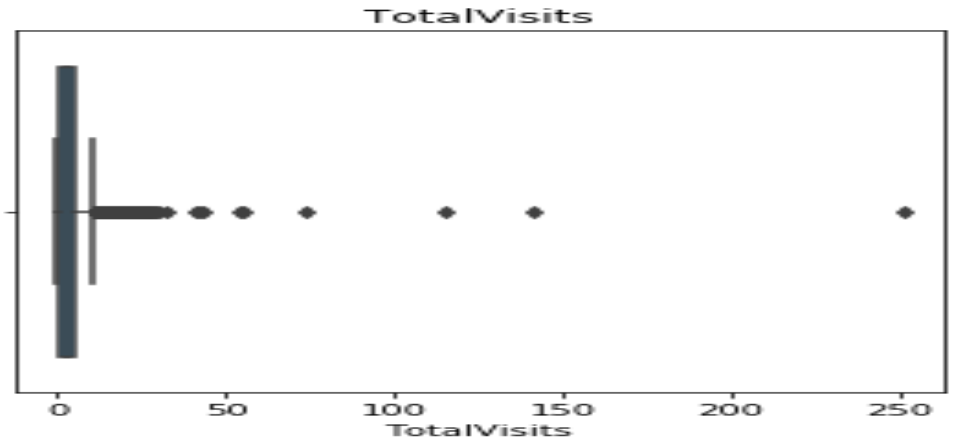
Summary

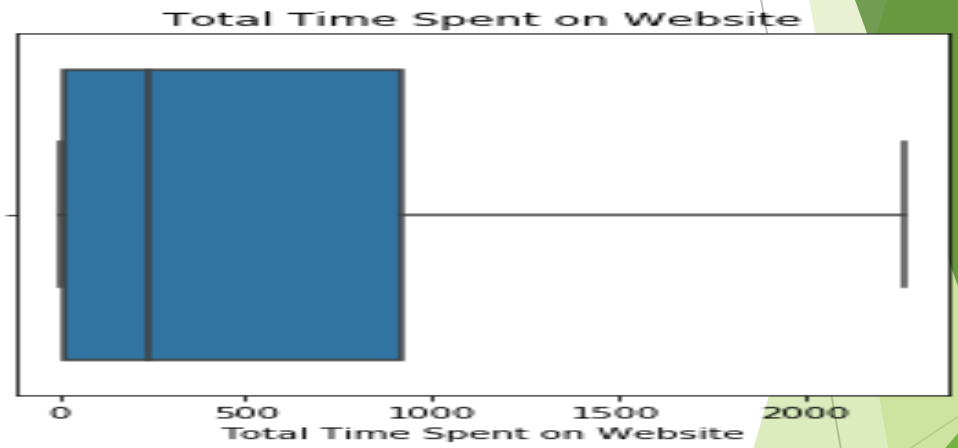
- Accuracy: 91%
- Recall: 82%
- Precision: 93%

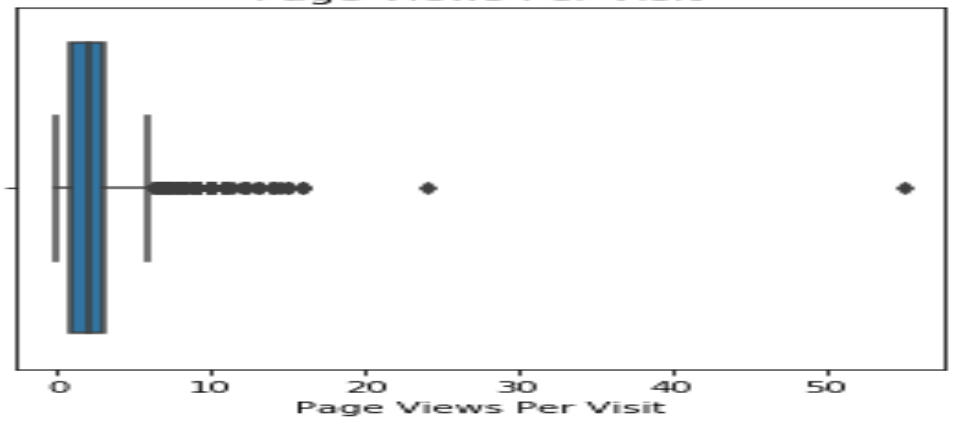
Exploratory Data Analysis & Outlier treatment

Finding the data imbalance: Clearly, there's imbalance in the data. Only 38% of the leads are converted and the remaining 62% are not converted.

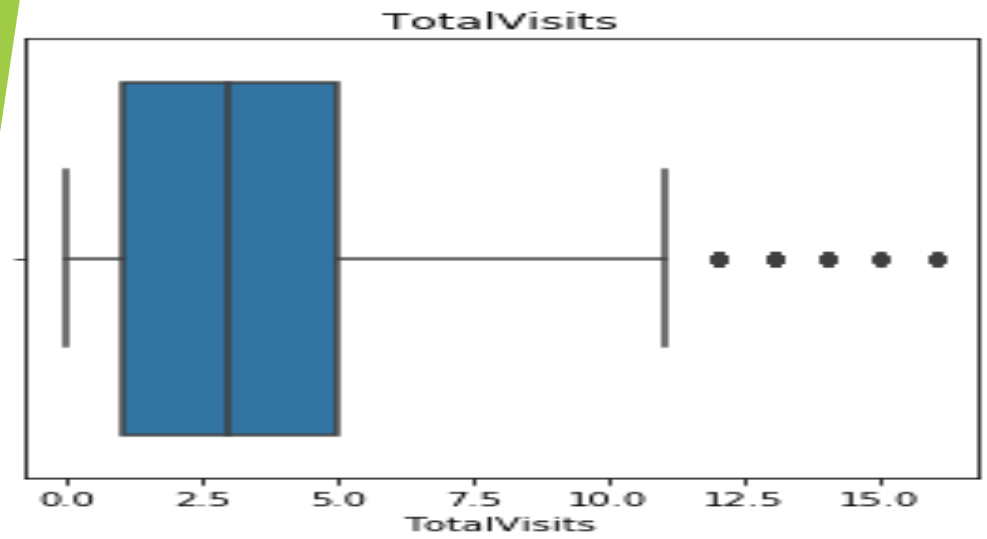
Visualizing numerical columns - Univariate Analysis

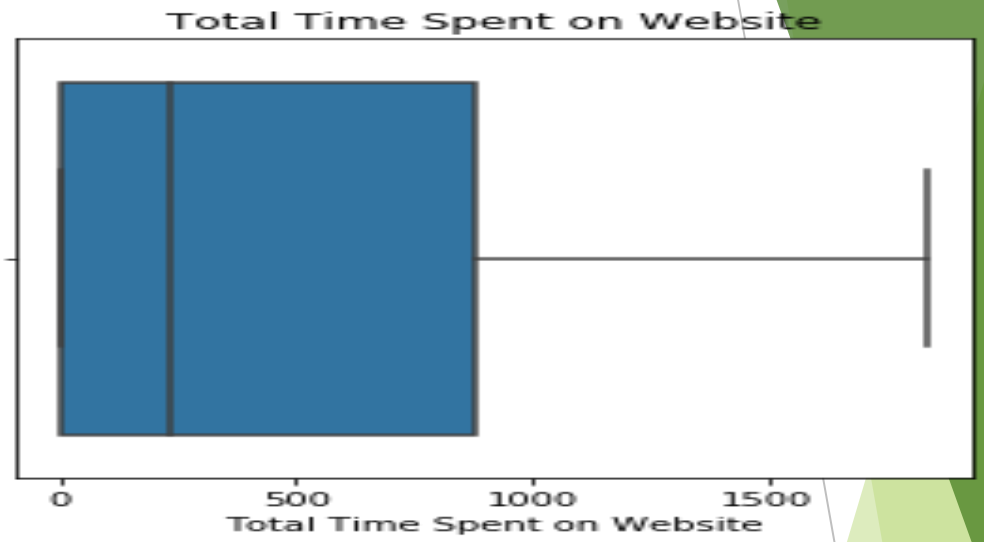


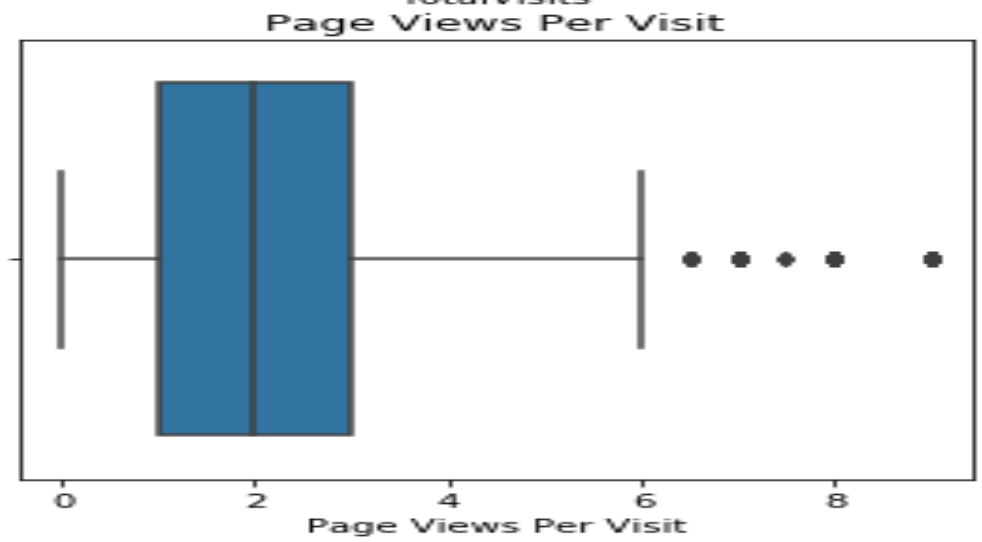




Handling Outliers: Removing outliers using IQR



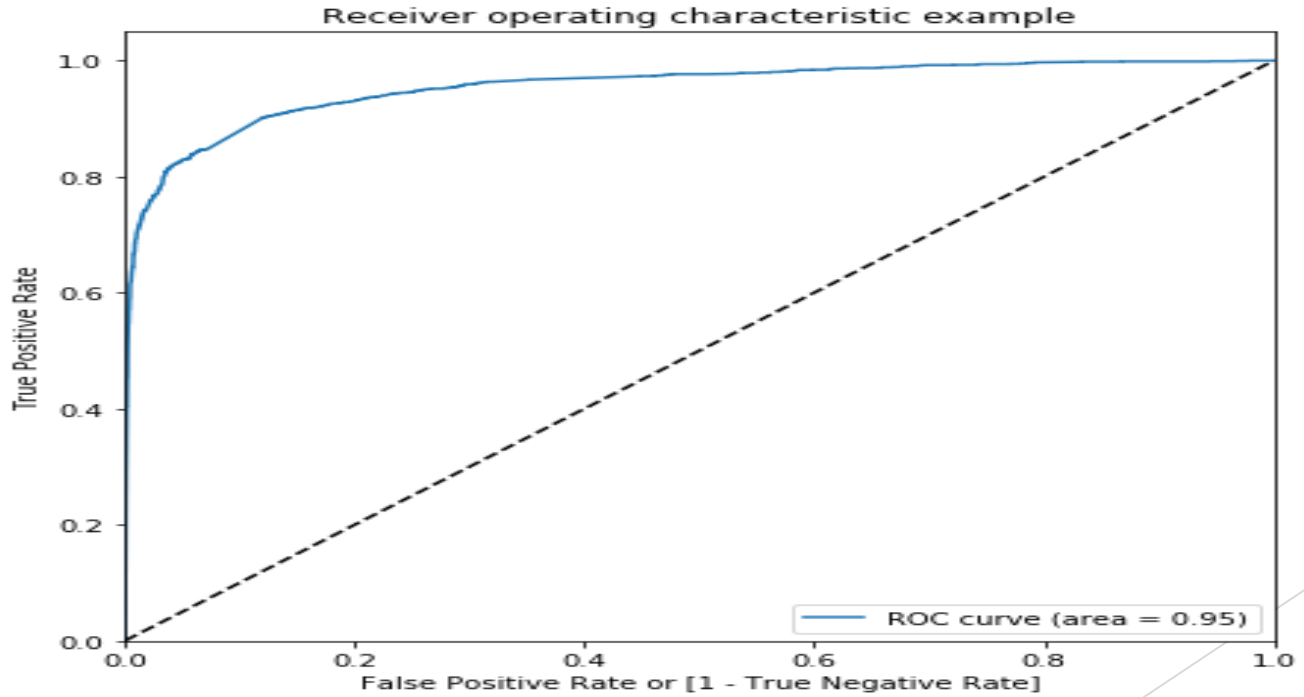




Plotting ROC Curve

After building the final model making prediction on it(on train set), we created ROC curve to find the model stability with auc score(area under the curve) As we can see from the graph plotted on the right side, the area score is 0.95 which is a great score.

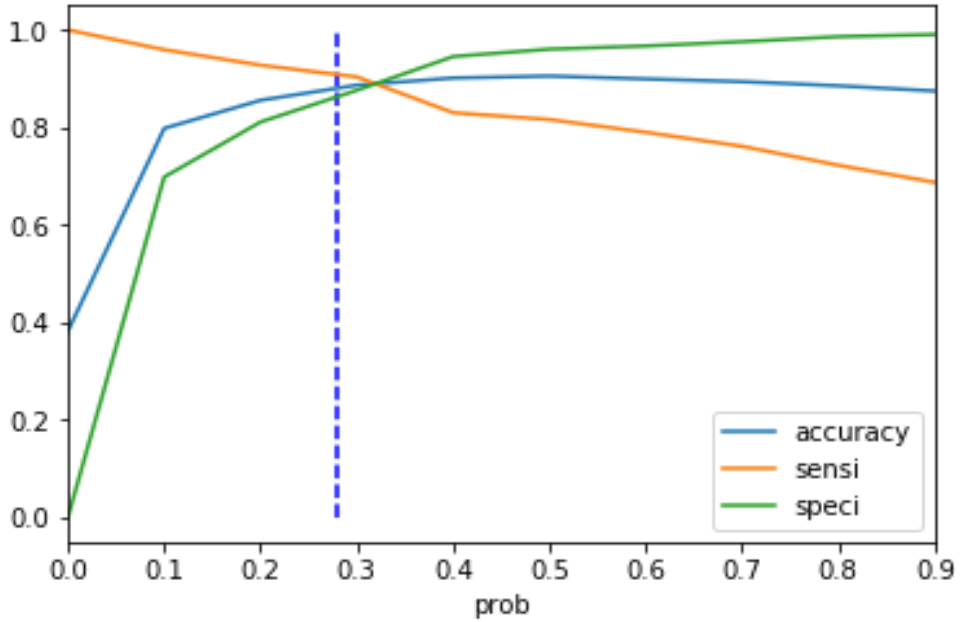
And our graph is leaned towards the left side of the border which means we have good accuracy.



Finding the optimal cutoff point

Now, we have created range of points for which we will find the accuracy, sensitivity and specificity for each points and analyze which point to chose for probability cutoff and we found that on 0.3 point all the score of accuracy, sensitivity and specificity are in a close range which is the ideal point to select and hence it was selected.

To verify our answer we plotted this in a graph – line plot which is on the right side and we stand corrected that the meeting point is close to 0.3 and hence we choose 0.3 as our optimal probability cutoff.



Precision and Recall

We used this cutoff point to create a new column in our final dataset for predicting the outcomes.

After this we did another type of evaluation which is by checking Precision and Recall

As we all know, Precision and Recall plays very important role in build our model more business oriented and it also tells how our model behaves.

Hence, we evaluated the precision and recall for this model and found the score as 0.73 for precision and 0.79 for recall.

Now, recall our business objective - the recall percentage I will consider more valuable because it is okay if our precision is little low which means less hot lead customers but we don't want to left out any hot leads which are willing to get converted hence our focus on this will be more on Recall than Precision.

i.e We get more relevant results - as many as hot lead customers from our model

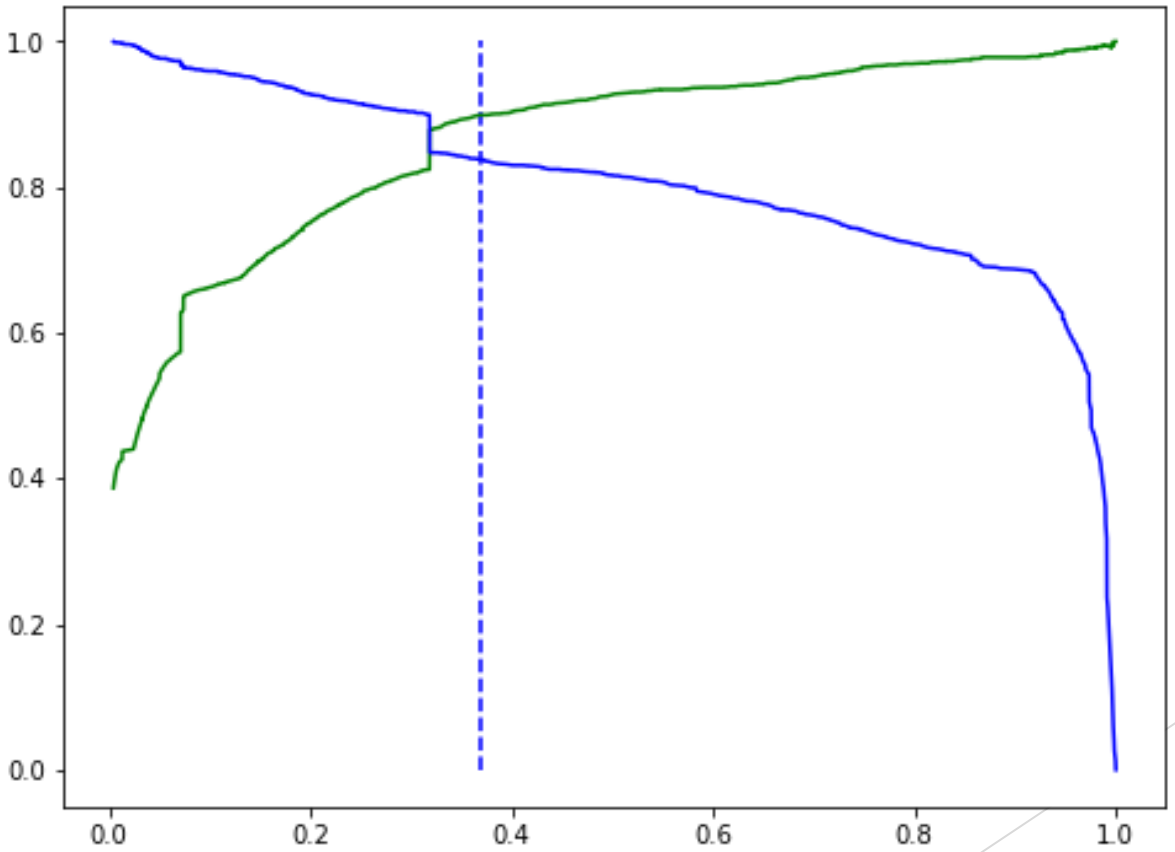
We created a graph which will show us the tradeoff between Precision and recall.

We found that there is a trade off between Precision and Recall and the meeting point is approximately at 0.38.

Precision and Recall trade-off

0.38 is tradeoff between Recall and Precision

Hence we can say that we can consider any prospect lead with conversion rate of 38% to become a Hot Lead



Final Observation

Train Data

- Accuracy : 90.53%
- Precision : 80%
- Recall : 91%
- F1-score : 85%

Test Data

Accuracy : 88.03%

- Precision : 91%
- Recall : 85%
- F1-score : 88%

Conclusion

To achieve maximum conversion with less amount of time spent on cold calling company should invest time on Hot Leads i.e, Leads who has more than 80% conversion rate. All the Potential Leads can be approached when the sales person have enough time to invest as the conversion rate is comparatively less.

Company should also focus on Lead Score which are greater than 80% to expedite the conversion rate

Thank you.