

Lead Scoring Assignment

- K Divakar Reddy
- B Praveen Reddy

Lead Scoring Case Study

- **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%.
- The company requires you to build a model wherein you need 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. The CEO has given a ballpark of the target lead conversion rate to be around 80%

Approach



Loading data and understanding the data by visualizing



Univariate Analysis



Dealing with missing values and Outlier treatment



Converting Categorical variables using encoding



Scaling the data



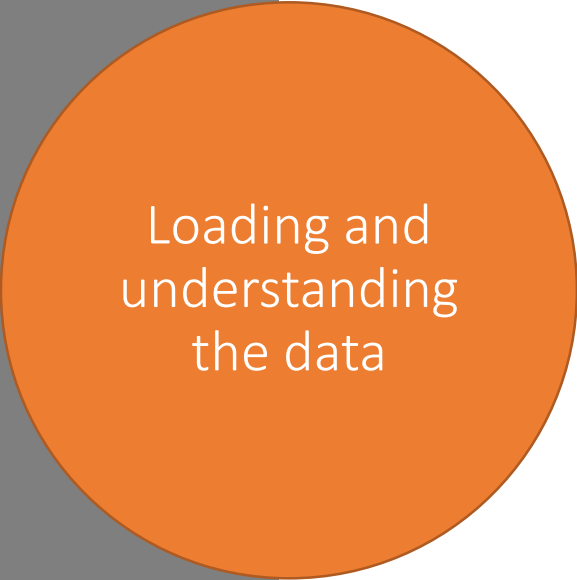
Model building (Logistic regression)



Model Evaluation



Inferences

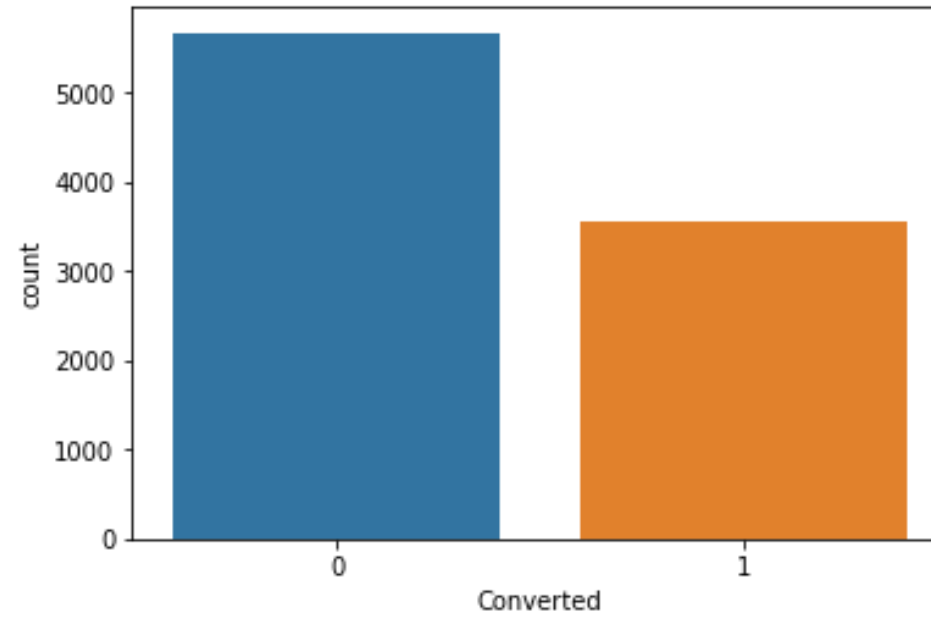
An orange circle with a thin black outline, positioned on the left side of the slide. It contains the text 'Loading and understanding the data' in white.

Loading and understanding the data

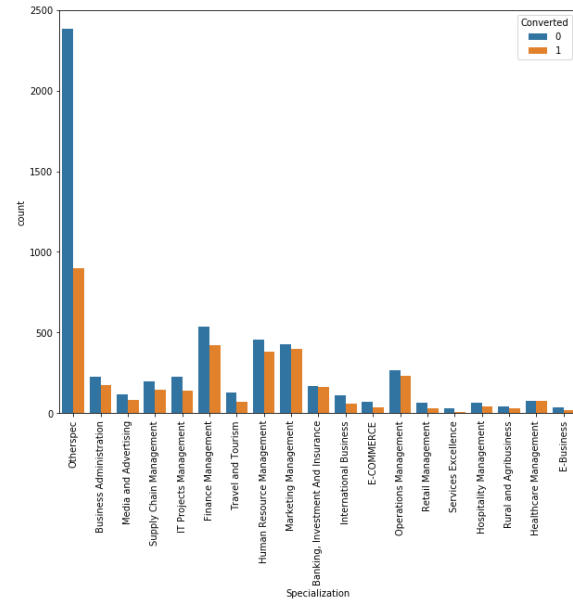
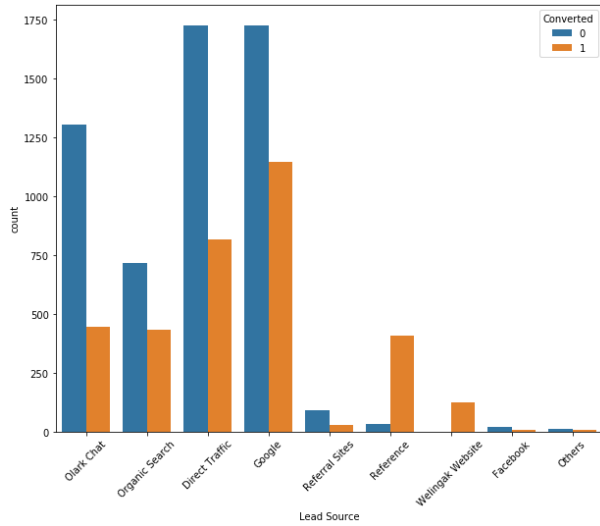
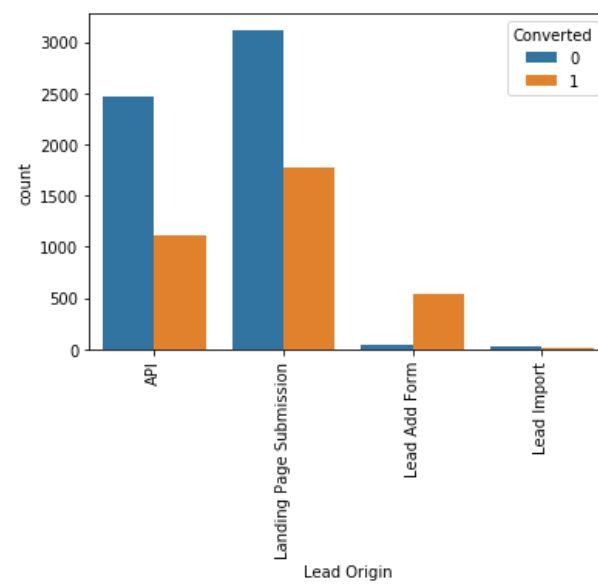
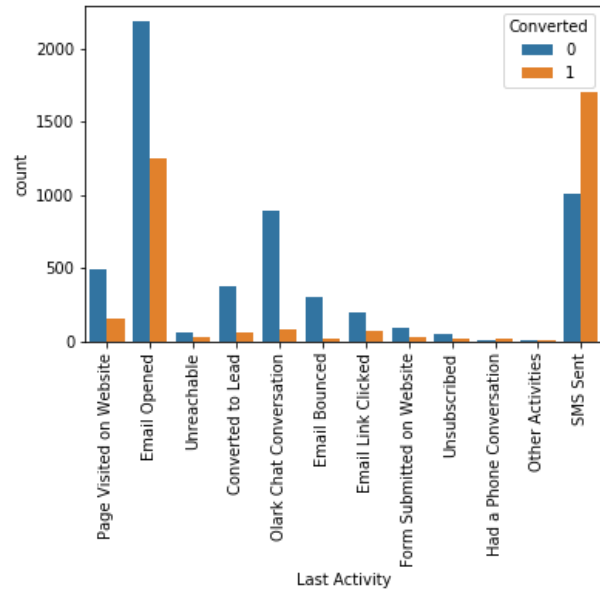
We noticed the shape of the data to be (9240,37) and observed missing values present in the data.

We observed the data set information using the functions `info()`, `describe()` and attribute `shape`.

Univariate Analysis



- From the above count plot it is visible that most of them are unconverted.

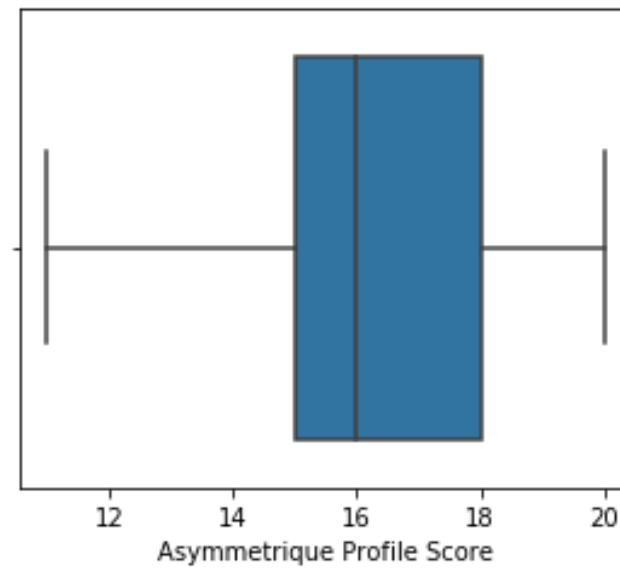
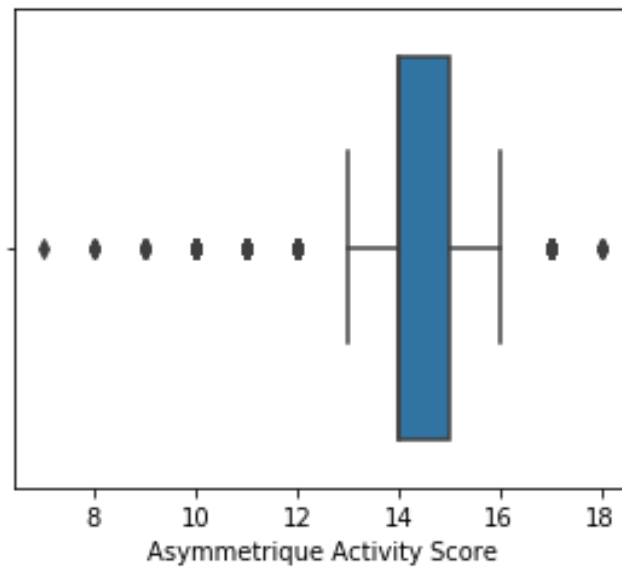
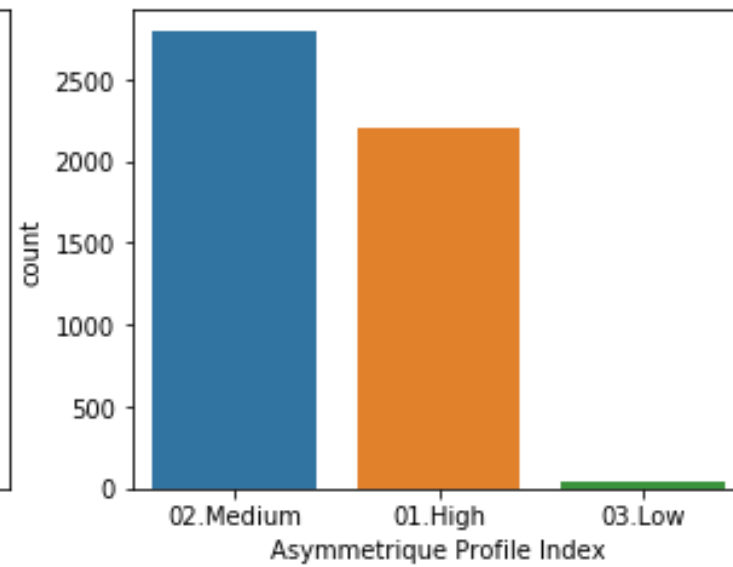
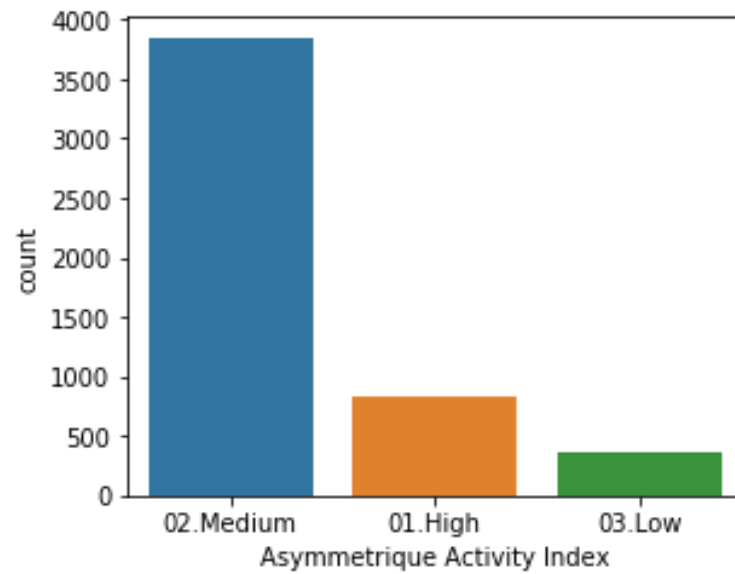


Segmentation analysis and inferences

- Major conversion happened in last activity are from SMS sent.
- Incase of Lead origin it is Landing page Submission.
- Incase of Lead source it is reference.
- lead conversion is bit high for Banking, Investment And Insurance ,Healthcare Management, Marketing Management specializations compare to others

Dealing with Missing values and outliers

- Firstly, we converted all the select values to Nan values and then we dropped the columns with more than 70% missing values. We also dropped the columns 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score' as they don't have any pattern and has 45% of missing values. From their boxplot it is clearly visible that they don't have any pattern. And we have imputed the remaining missing column values with mode of the column. Finally we did not lose significant amount of data and are left 98% of data which is good to go.



Converting Categorical variables using encoding and Scaling



We treated the categorical columns using dummy variables.



We performed label encoding technique on the variables and mapped “Yes/no” columns with 1 and 0.



We split the data to train(70%) and test(30%) data, then performed scaling using StandardScaler.

Model Building

- As we have so many features to work with, first we will eliminate some of them by using RFE technique and finally left took 30 variables. Later we imported statsmodel api and built a model with these 30 variables. We analyzed the summary taking p values into consideration. And we also analyzed the VIF values of these columns. We started eliminating the variables manually with high p and high VIF one by one, then rebuilt the model. This process is carried out until the p value is less than 0.05 and VIF less than 5. Finally we arrived at a model with desired p and VIF values.

Final Model
with desired
p values and
VIF values.

	Coef	Std err	Z	P> Z	0.025	0.975
const	0.4521	0.338	1.336	0.182	-0.211	1.116
Do Not Email	-1.0497	0.247	-4.251	0.000	-1.534	-0.566
Total Time Spent on Website	1.1438	0.063	18.233	0.000	1.021	1.267
Direct Traffic	-1.7786	0.174	-10.248	0.000	-2.119	-1.438
Google	-1.3925	0.161	-8.671	0.000	-1.707	-1.078
Organic Search	-1.4360	0.199	-7.221	0.000	-1.826	-1.046
Referral Sites	-1.4696	0.484	-3.039	0.002	-2.417	-0.522
Welingak Website	4.9795	1.024	4.862	0.000	2.972	6.987
Last Activity_SMS Sent	1.9435	0.118	16.462	0.000	1.712	2.175
What is your current occupation_Unemployed	-0.7280	0.317	-2.297	0.022	-1.349	-0.107
Tags_Already a student	-3.3801	1.040	-3.250	0.001	-5.418	-1.342
Tags_Busy	0.9498	0.237	4.002	0.000	0.485	1.415
Tags_Closed by Horizon	6.8522	0.730	9.392	0.000	5.422	8.282
Tags_Interested in full time MBA	-1.7148	0.783	-2.190	0.029	-3.250	-0.180
Tags_Interested in other courses	-1.5704	0.360	-4.359	0.000	-2.277	-0.864
Tags_Lost to EINS	7.0277	0.819	8.580	0.000	5.422	8.633
Tags_Ringing	-3.2705	0.240	-13.617	0.000	-3.741	-2.800
Tags_Will revert after reading the email	4.6382	0.198	23.443	0.000	4.250	5.026
Tags_switched off	-3.6609	0.615	-5.957	0.000	-4.865	-2.456
Lead Quality_Worst	-2.2672	0.676	-3.355	0.001	-3.592	-0.943
Last Notable Activity_Email Link Clicked	-1.2361	0.439	-2.813	0.005	-2.097	-0.375
Last Notable Activity_Modified	-1.7128	0.129	-13.314	0.000	-1.965	-1.461
Last Notable Activity_Olark Chat Conversation	-1.2458	0.372	-3.351	0.001	-1.974	-0.517

VIF values for final model

- 4 Google 2.03
- 3 Direct Traffic 1.99
- 17 Tags_Will revert after reading the email 1.64
- 5 Organic Search 1.58
- 10 Tags_Already a student 1.54
- 19 Lead Quality_Worst 1.53
- 2 Total Time Spent on Website 1.47
- 21 Last Notable Activity_Modified 1.24
- 16 Tags_Ringing 1.23
- 9 What is your current occupation_Unemployed 1.21
- 8 Last Activity_SMS Sent 1.19
- 14 Tags_Interested in other courses 1.15
- 12 Tags_Closed by Horizzon 1.10
- 22 Last Notable Activity_Olark Chat Conversation 1.07
- 1 Do Not Email 1.07
- 7 Welingak Website 1.07
- 6 Referral Sites 1.07
- 18 Tags_switched off 1.06
- 11 Tags_Busy 1.06
- 15 Tags_Lost to EINS 1.05
- 13 Tags_Interested in full time MBA 1.04
- 20 Last Notable Activity_Email Link Clicked 1.04

Model Evaluation

- We made prediction after building the final model and calculated the stats parameters such as sensitivity, specificity, accuracy.
- Confusion matrix with cutoff 0.5 comes out to be as

`[[3811, 166],[297, 2077]]`

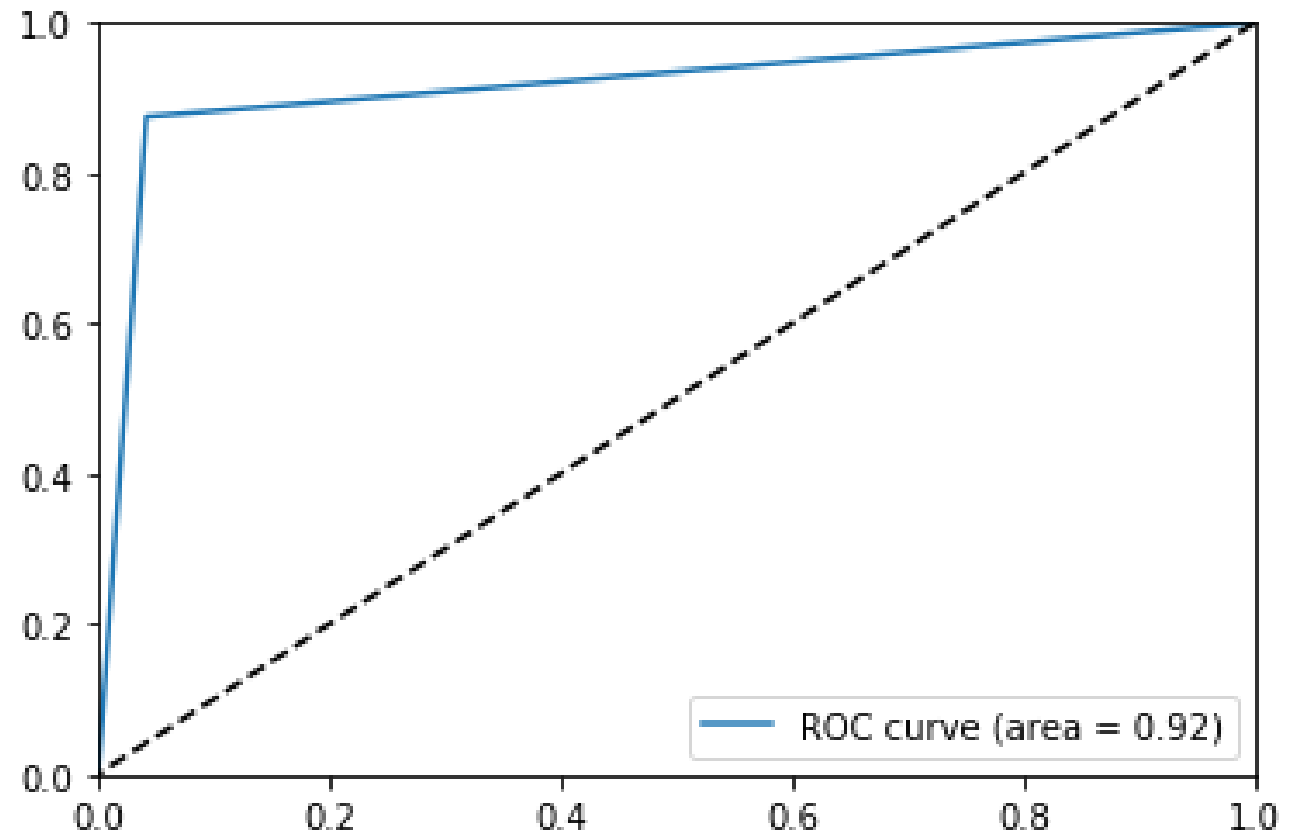
Accuracy score = 0.92

ROC curve with area under curve value as 0.92.

ROC curve

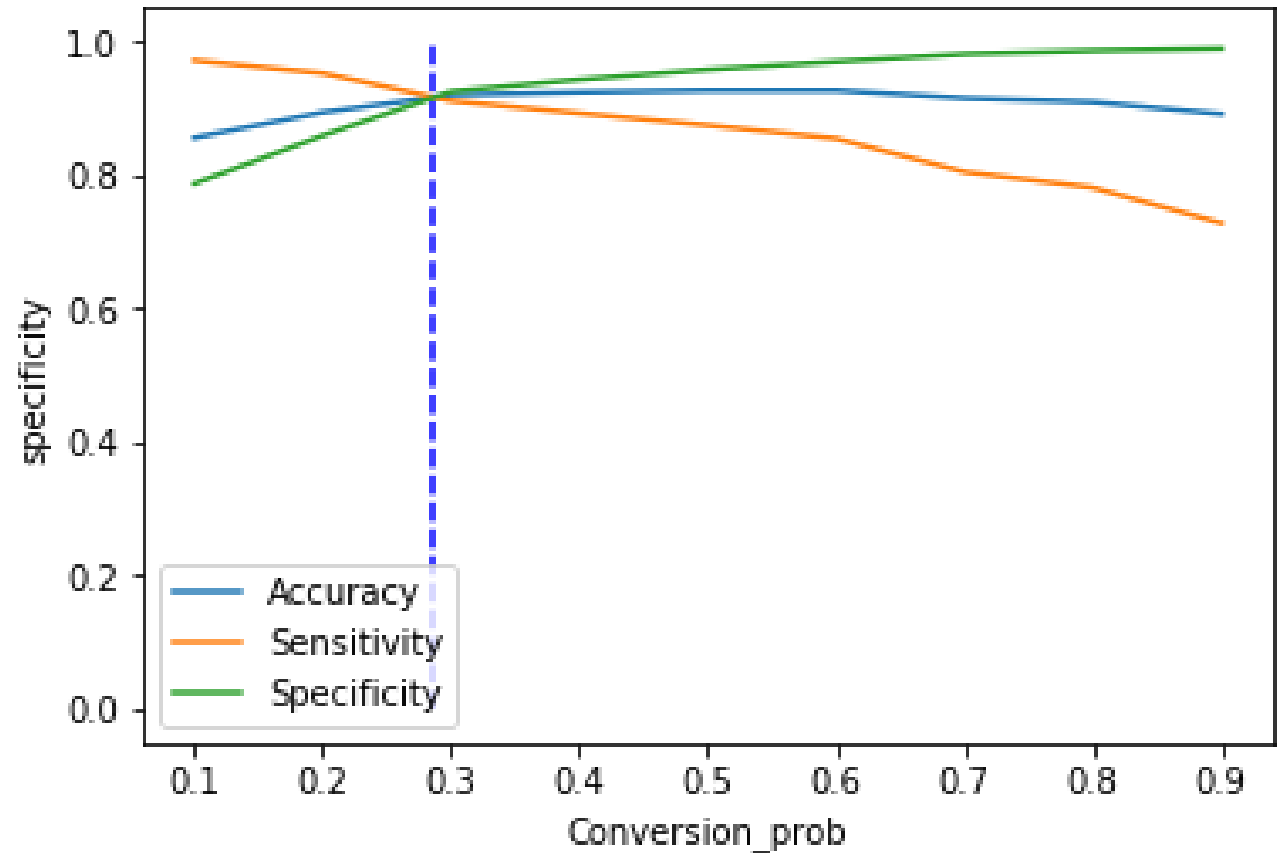
From the graph it is observed that the area under curve comes out be 0.92 which is very good indication.

It almost resembles as right angle triangle.



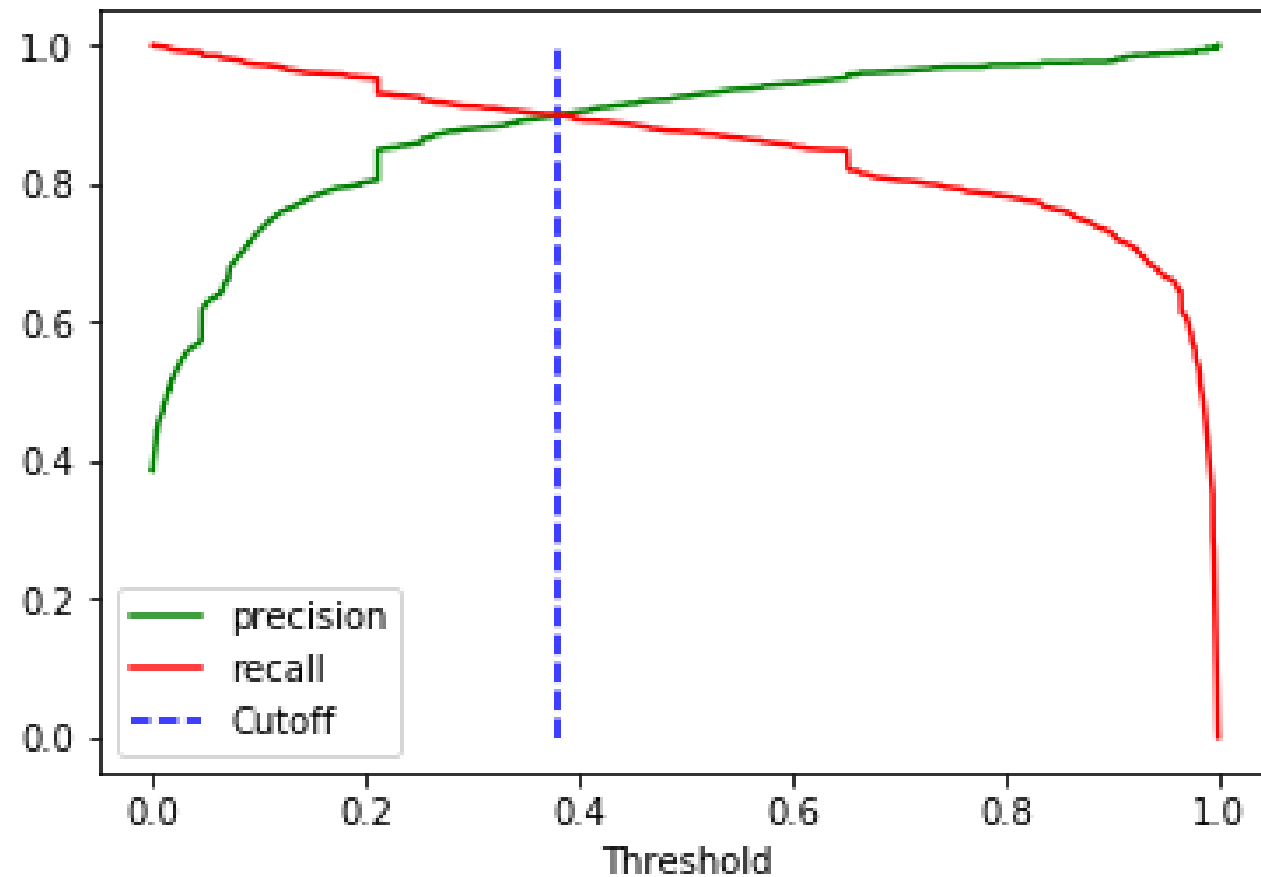
Sensitivity and specificity plot

- From the curve 0.285 is the optimal point where we have sensitivity and specificity are high. Hence we chose the point as 0.285 and made predictions again which resulted in 0.92 accuracy.
- Sensitivity= 0.91
- specificity= 0.92



Precision recall plot

- Recall score=0.914
- Precision score= 0.876



Predictions on test data

- Finally, we predicted the test data on the trained model which resulted in the following parameters.
- Sensitivity = 0.92
- Specificity = 0.918
- Accuracy score = 0.919
- While we checked both Sensitivity and specificity as well as precision and recall metrics, we have considered the optimal cut off based on specificity and sensitivity for calculating the final prediction.

Inferences

To improve conversion rate or to reach goal of 80% conversion. We should concentrate on the following leads.

- Leads identification are high for API and landing Page Submission but low conversion rate.
- Lead source is high from Direct Traffic, Google, Olark Chat, Organic Search but conversion rate is low (<40%).
- Last activities like Email opened and SMS sent have more lead but less conversion. So, try to have phone conversion with such leads. Since phone conversions had good conversion rate.
- Most of the leads are unemployed and they have low conversion rate. To increase conversion rate, provide some scholarship schemes or paid internships to them to cut down their fee expenses.
- Lead who spend more time in website have good conversion rate. So, ask leads to visit site and allow to see them some demo lectures and how internship program is designed videos
- Concentrate on leads from API, landing page submissions who has low conversion rates. Try to increase the leads from lead add forms since they have high conversion rates.
- Top variables that improve probability of lead getting converted are
 - Tags_Lost to EINS
 - Tags_Closed by Horizon
 - Welingak Website
 - Tags_Will revert after reading the email
 - Last Activity_SMS Sent
 - Total Time Spent on Website



Thank you