Lead Scoring Case Study

IDENTIFICATION OF HOT LEADS THAT ARE MOST LIKELY TO CONVERT INTO PAYING CUSTOMERS

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites 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.
- 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
- Typical lead conversion rate at X education is around 30%.
- Company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- Sales team can then focus on these ‘Hot Leads’ for bettering the conversion rate

Objective

- Categorize the leads and assign a lead score to each of the leads such that the customers with higher lead score will become hot leads.
- Target lead conversion rate to be 80%.
- For the above, build a logistic regression model

Road Map

In this case study as a analyst we need to build a model so that we can distinguish to active lists from a set of dataset so that we can get a way-out to for maximum conversion of lead . To accomplish this we need to follow below mentioned road map

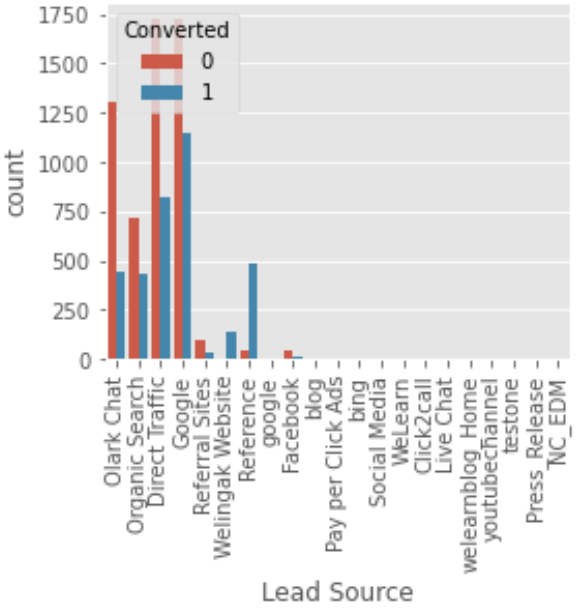
- Importing Data
- Dataframe Inspections
- Data Preparation (Encoding Categorical Variables, Handling Null Values)
- EDA (univariate analysis, outlier detection, checking data imbalance)
- Dummy Variable Creation
- Test-Train Split
- Feature Scaling
- Looking at Correlations
- Model Building (Feature Selection Using RFE, Improvising the model further inspecting adjusted R-squared, VIF and p-vales)
- Build final model
- Model evaluation with different metrics Sensitivity, Specificity.

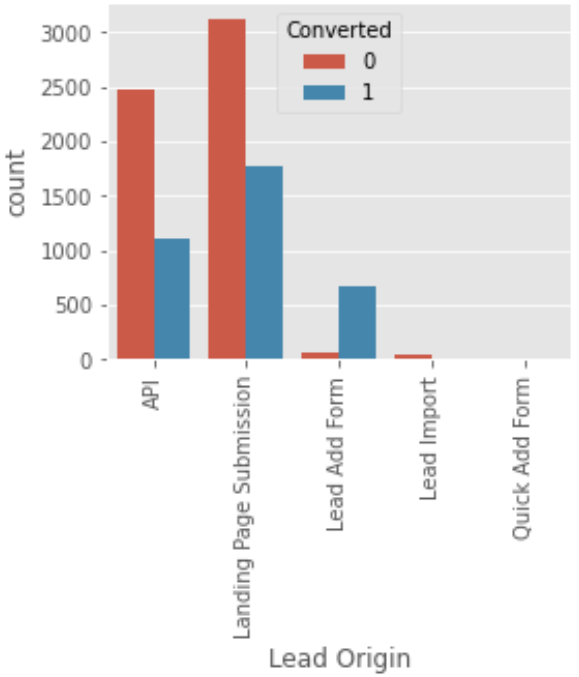
EDA

UNIVARIATE ANALYSIS, OUTLIER DETECTION, CHECKING DATA IMBALANCE

API and Landing Page Submission has less conversion rate(~30%) but counts of the leads from them are considerable The count of leads from the Lead Add Form is pretty low but the conversion rate is very high Lead Import has very less count as well as conversion rate and hence can be ignored

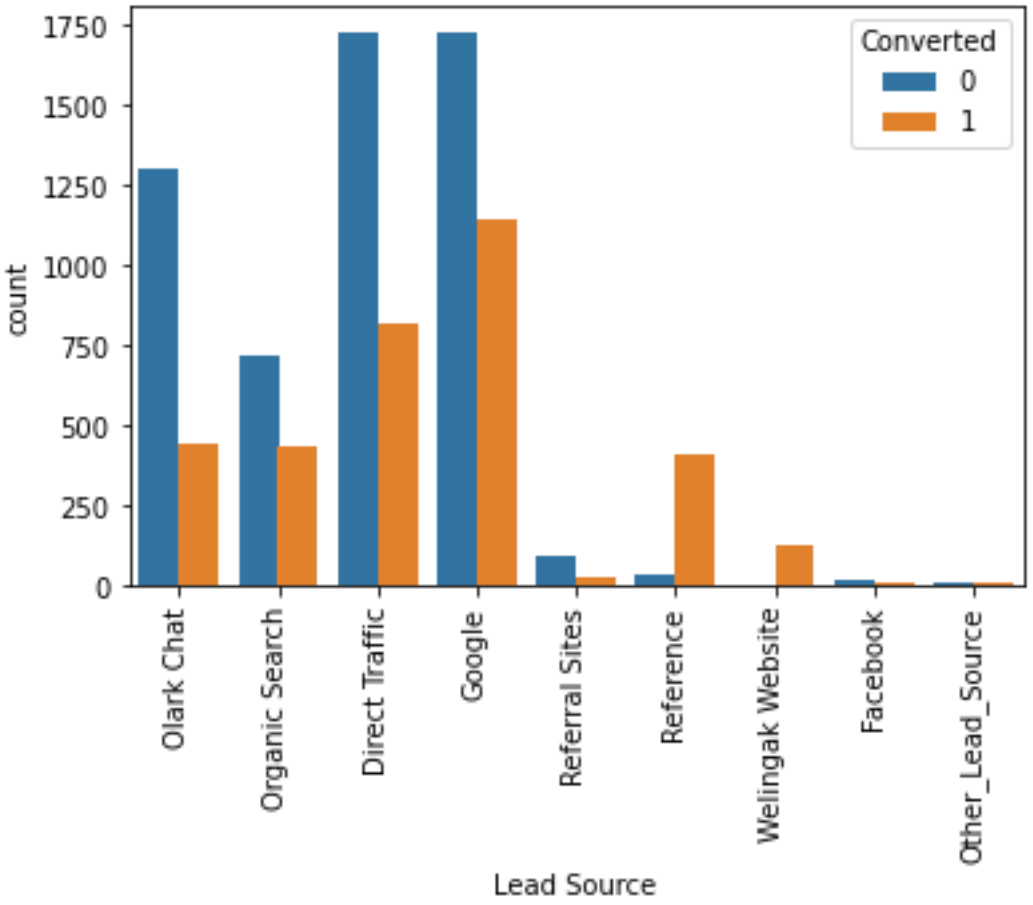
To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' and also increasing the number of leads from 'Lead Add Form





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

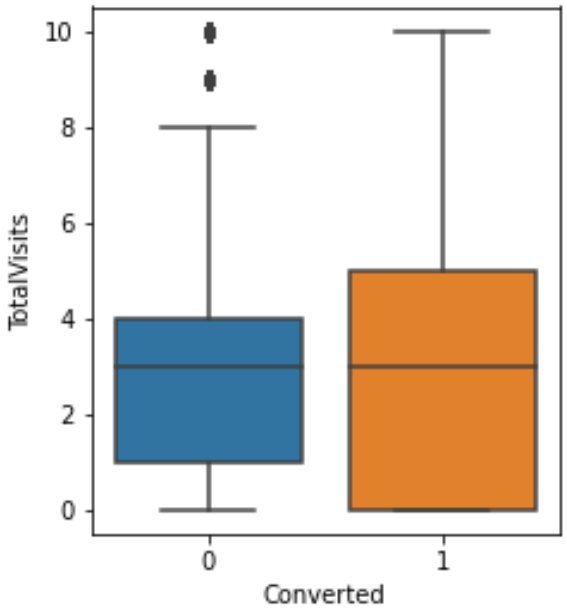
To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website'

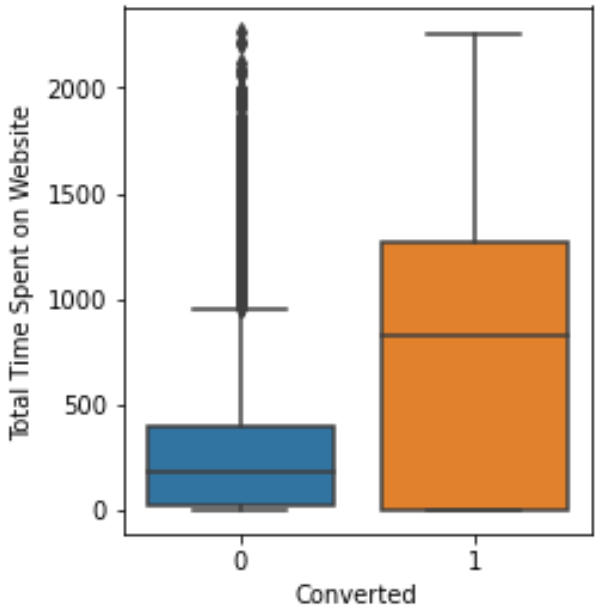


The median of both the conversion and non-conversion are same and hence nothing conclusive can be said using this information

Users spending more time on the website are more likely to get converted

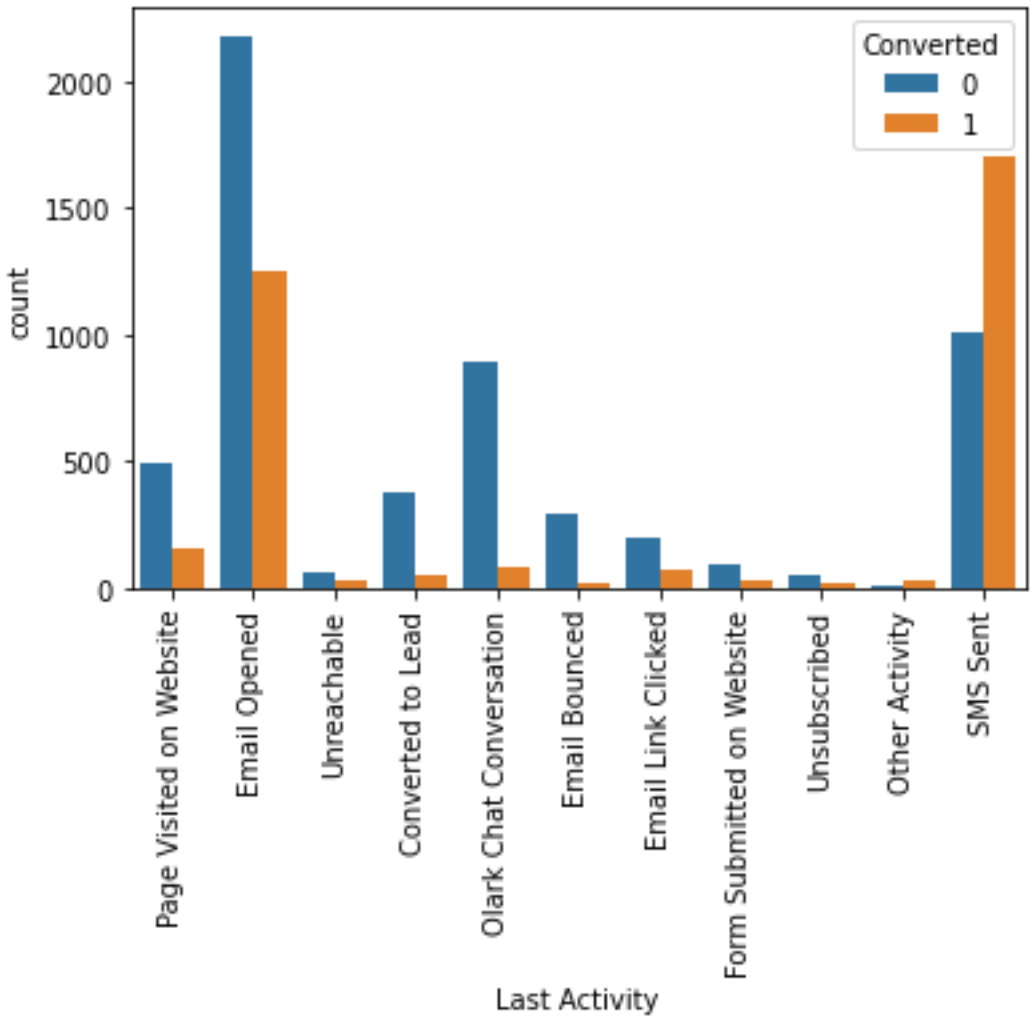
Websites can be made more appealing so as to increase the time of the Users on websites





The count of lst activity as "Email Opened" is max The conversion rate of SMS sent as last activity is maximum

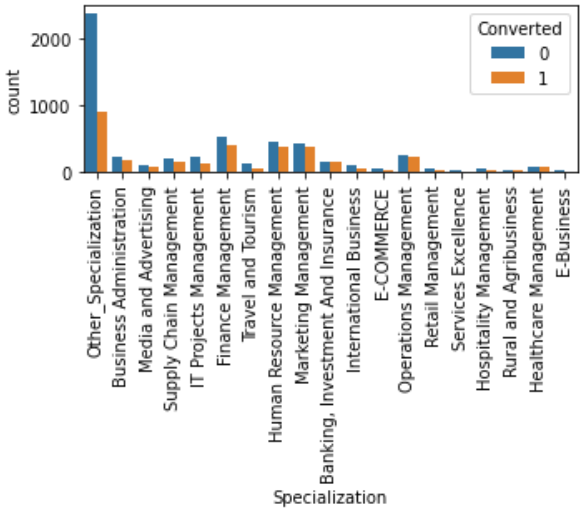
We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sen.

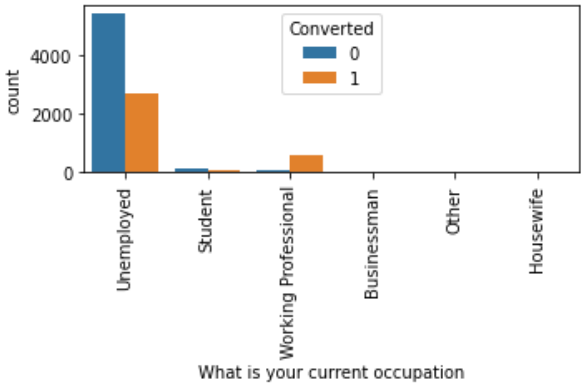


Looking at above plot, no particular inference can be made for Specialization Looking at above plot, we can say that working professionals have high conversion rate Number of Unemployed leads are more than any other category

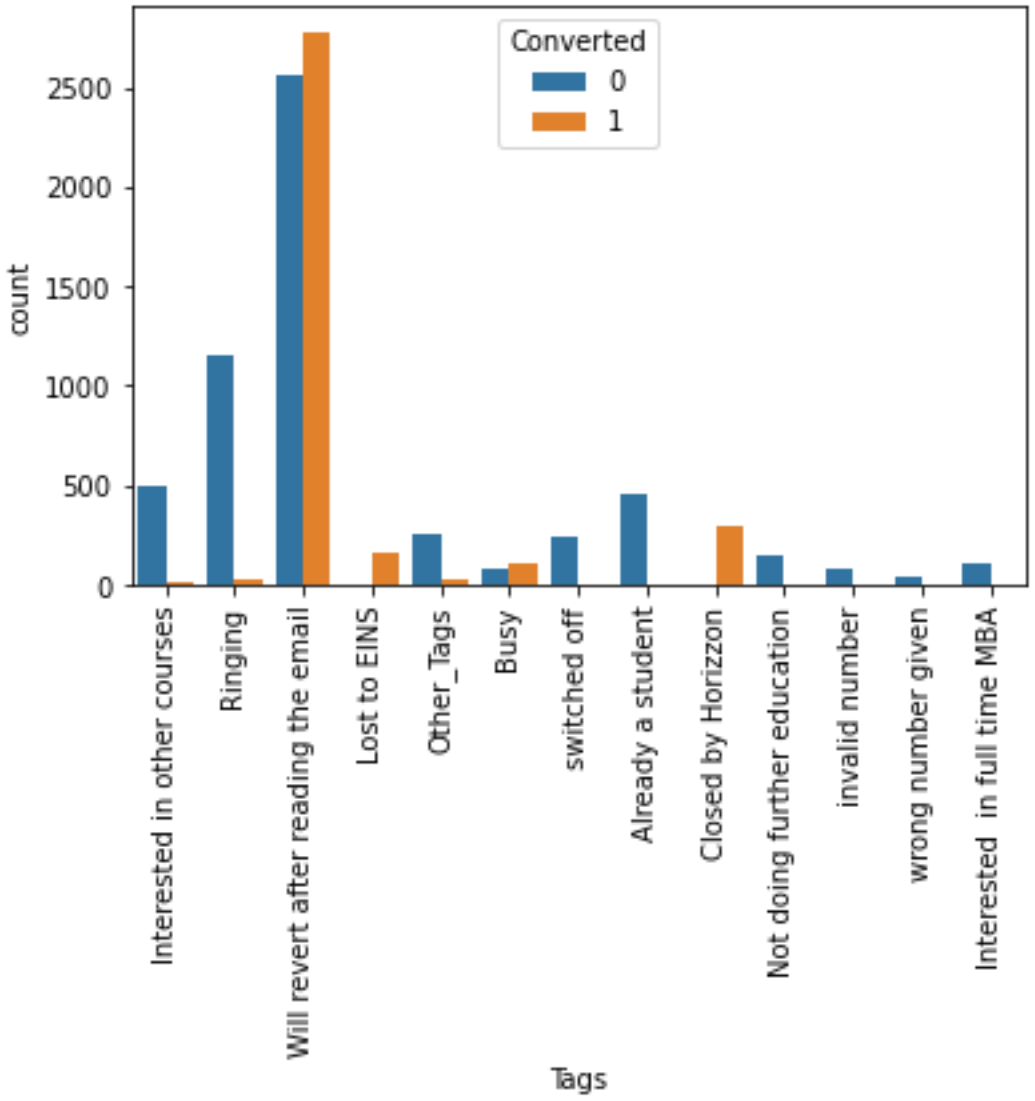
To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads

Country, What matters most to you in choosing a course, City columns have most values corresponding to one value such as India for Country, Mumbai for city and hence there is no particular insights for these columns





'Will revert after reading the email' and 'Closed by Horizzon' have high conversion rate

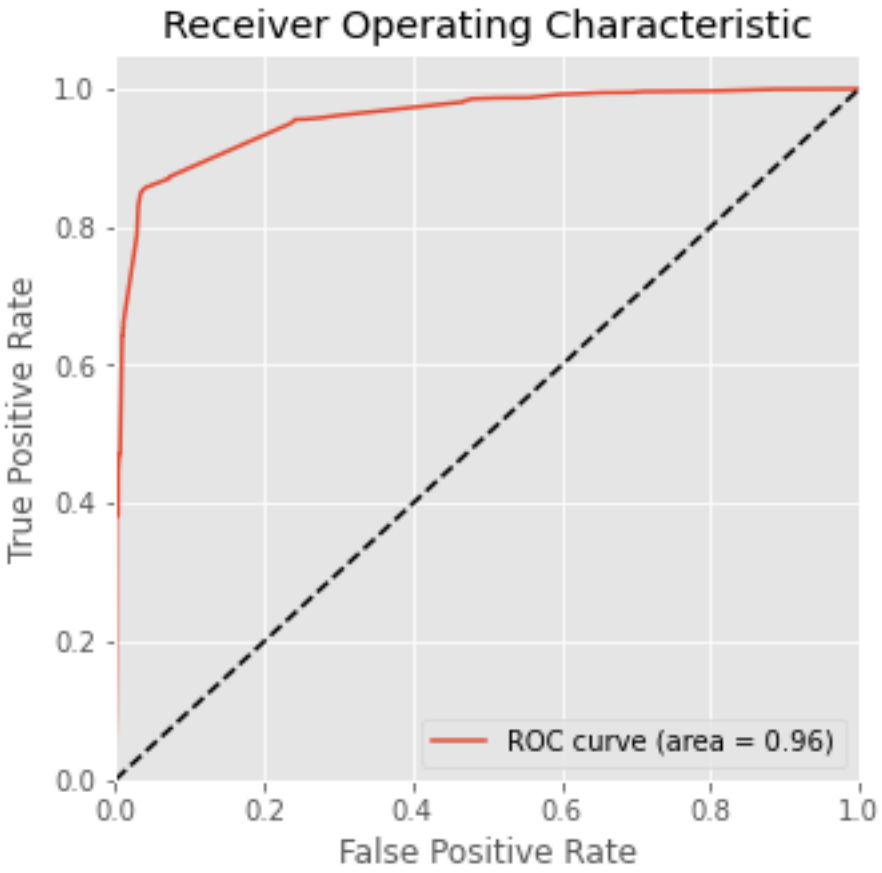


Summary

1. To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' Lead Origins and also increasing the number of leads from 'Lead Add Form'
2. To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'Google', 'Olark Chat', 'Organic Search', 'Direct Traffic' and also increasing the number of leads from 'Reference' and 'Welingak Website‘
3. Websites can be made more appealing so as to increase the time of the Users on websites
4. We should focus on increasing the conversion rate of those having last activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent
5. To increase overall conversion rate, we need to increase the number of Working Professional leads by reaching out to them through different social sites such as LinkedIn etc. and also on increasing the conversion rate of Unemployed leads
6. We also observed that there are multiple columns which contains data of a single value only. As these columns do not contribute towards any inference, we can remove them from further analysis

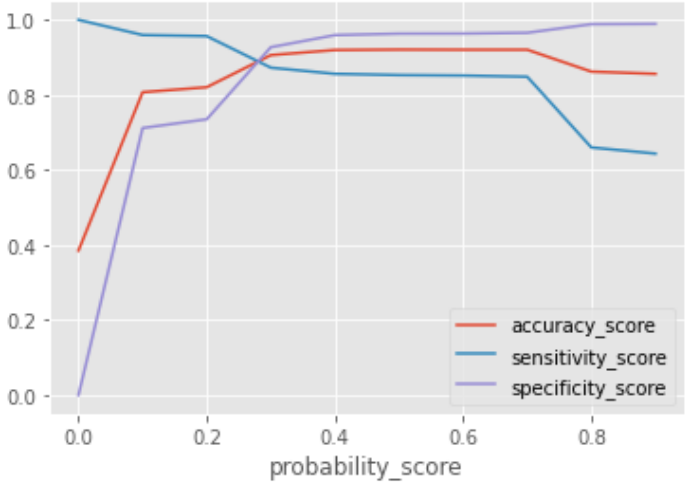
ROC CURVE

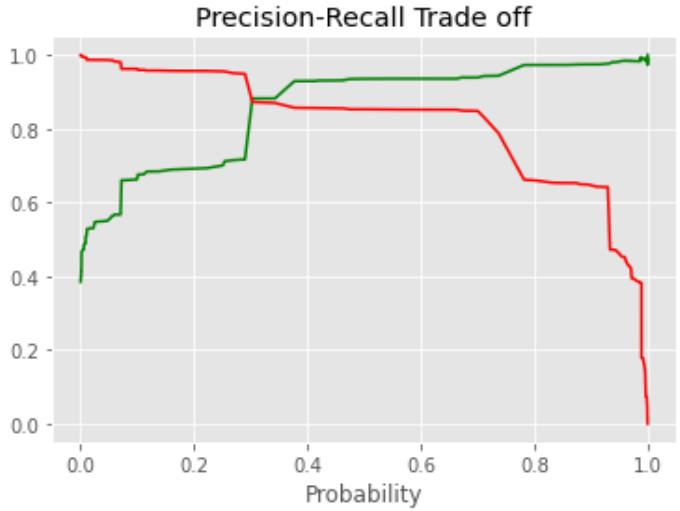
The ROC curve looks acceptable. 2. Area under the curve = 0.96



In Sensitivity-Specificity-Accuracy plot 0.27 probability looks optimal. In Precision-Recall Curve 0.3 looks optimal.

We are taking 0.27 is the optimum point as a cutoff probability and assigning Lead Score in training data.





Model Evaluation

Train Data-Confusion Matrix

Predicted Actual	Not converted	Converted
Not converted	2987	918
Converted	124	2322

Accuracy	83.59%
Precision	71.6%
Sensitivity	94.9%
Specificity	76.5%

Model Prediction

Test Data-Confusion Matrix

Predicted Actual	Not converted	Converted
Not converted	1303	431
Converted	71	918

Accuracy	81.5%
Precision	68.0%
Sensitivity	92.8%
Specificity	75.1%

Conclusion

1. The logistic regression model is used to predict the probability of conversion of a customer.
2. Optimum cut off is chosen to be 0.27 i.e. any lead with greater than 0.27 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.27 or less probability of converting is predicted as Cold Lead (customer will not convert)
3. The top three categorical/dummy variables in the final model are 'Tags_Lost to EINS', 'Tags Closed by Horizzon', 'Lead Quality_Worst' with respect to the absolute value of their coefficient factors
    - Tags_Lost to EINS (Coefficient factor = 9.578632)
    - Tags_Closed by Horizzon (Coefficient factor = 8.555901)
    - Lead Quality_Worst (Coefficient factor =-3.943680)
4. The final model has Sensitivity of 0.928, this means the model is able to predict 92% customers out of all the converted customers, (Positive conversion) correctly.
5. The final model has Precision of 0.68, this means 68% of predicted hot leads are True Hot Leads.

Thank You