Lead Scoring Casestudy

Presented by

- ❑ Pradhan Nayak
- ❑ Vijay Kumar Singh

Problem Statement and Business Goals

- An education company named X Education sells online courses to industry professionals.

- 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%.

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

- 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach to solve the case study

1.Data cleaning and inspection: check for duplicate values, check for null values, drop columns with higher percentage null values, imputing and handling 'nan' values.
2.EDA: Univariate and Bivariate Analysis: perform count plot and correlation matrix to check. the distribution and multicollinearity of the data.
3.Outlier: perform soft capping of numeric columns to remove outliers.
4.Data preparation: Binary mapping of categorical variables and dummy variable creation.
5. Perform Train-Test split: create a target variable and independent variables.
5.FeatureScaling: scaling the data with standardscaler preprocessor.
6.Model Building: Running first train model.
7.Feature selection using RFE and assessing the model with stats model.
8. Checking p values, VIF and confusion matrix.
9.Calculating metrics and plotting ROC curve.
10.Finding optimal cut-off point and assign lead score.
11. Calculate precision recall tradeoff.
12. Perform predictions on the Test data set.

Data Manipulations performed.

1. 7 columns have been dropped with 40% or more having null values.

2. Total number of rows : 9103 and columns: 15 used for model Building.

3. Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque columns have single value features.

4. 15 columns have been dropped because they don’t add value to analyze our model or because the data in those columns are highly skewed.

Univariate analysis of categorical variables.

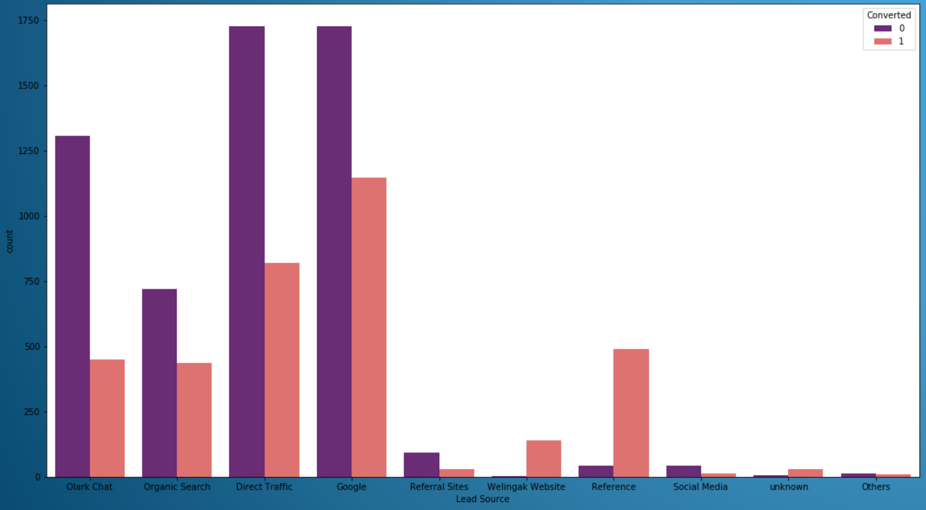
Points to be concluded from the graph.

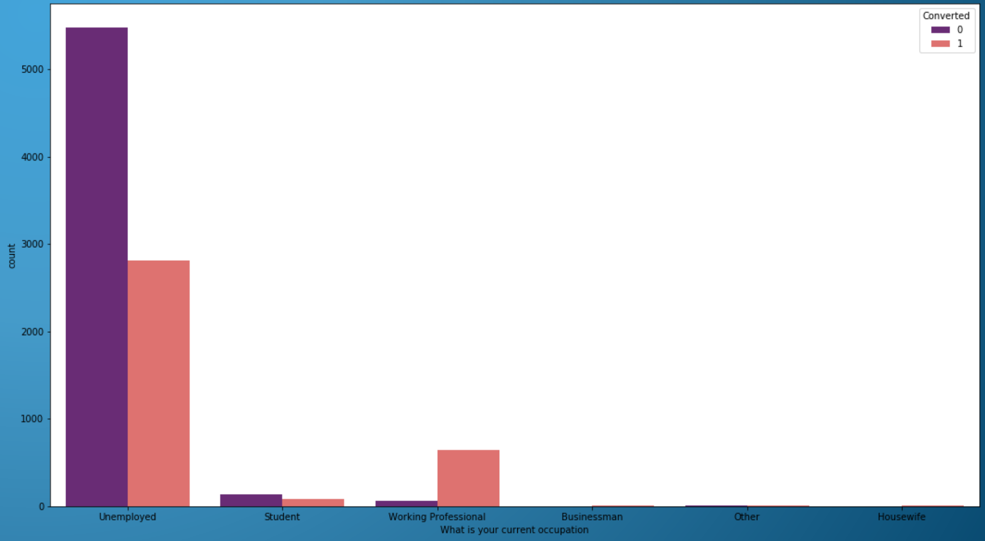
What is your occupation:

1. customers who are Unemployed are the ones who enquire more about X education online.
2. working professionals tend to get converted more.

Lead source

1. google and direct traffic seems to have high number of leads
2. wellingak website and reference tends to have a high conversion rate.





Univariate analysis of categorical variables.

Points to be concluded from the graph.

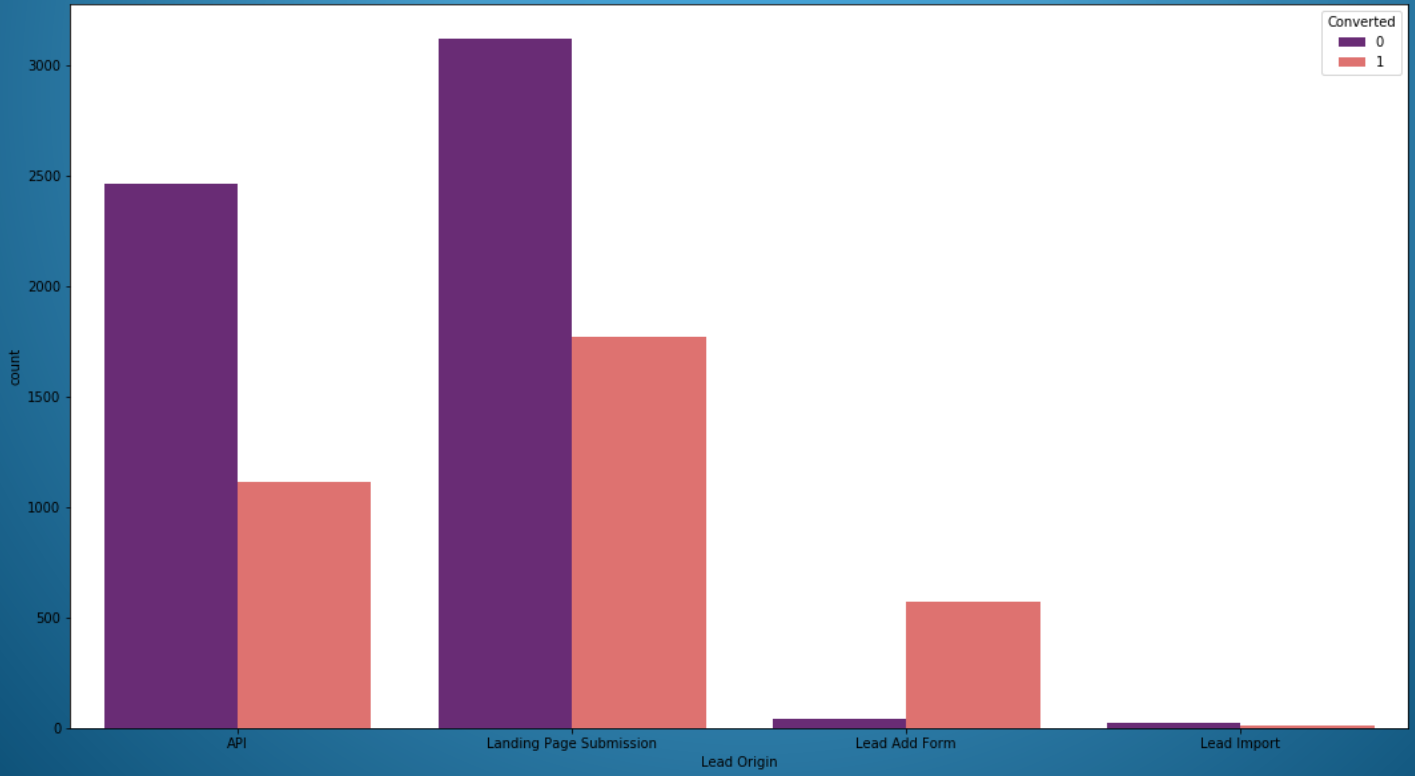
Lead origin

1.API and Landing page submission tend to have higher numbers and better lead conversion rate.

2.Lead Add form has a higher lead conversion rate but lesser in numbers.

3.lead import tend to have the lowest number and less lead conversion.

4. in order to increase the conversion rate company should focus more on getting leads from API and landing page submissions and emphasis to generate more lead conversions from lead add form.



Univariate analysis of categorical variables.

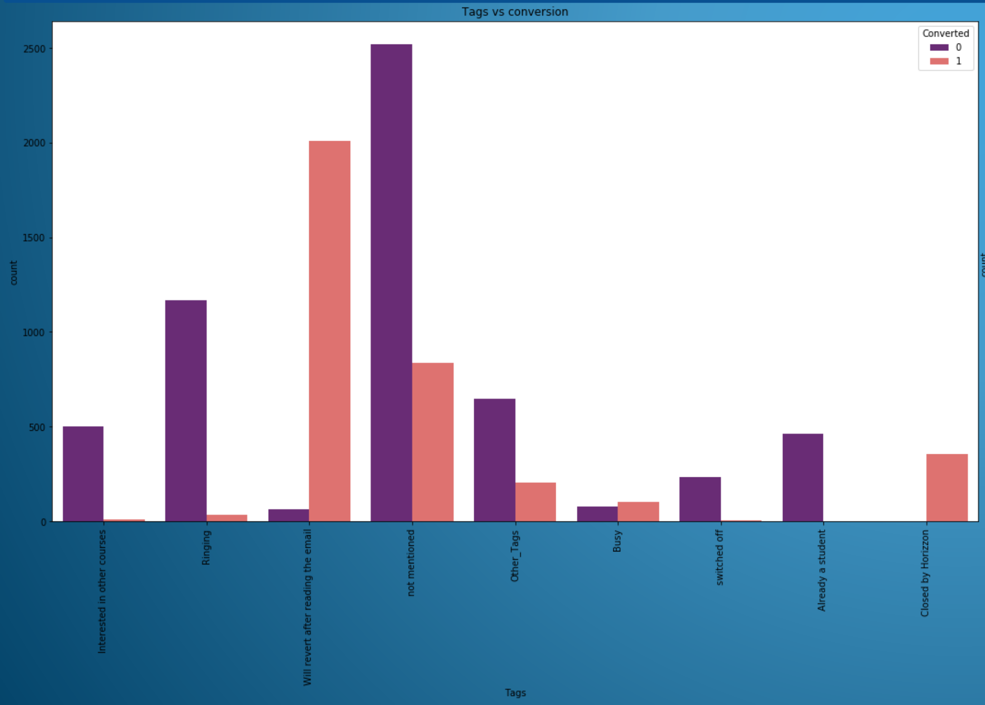
Points to be concluded from the graph.

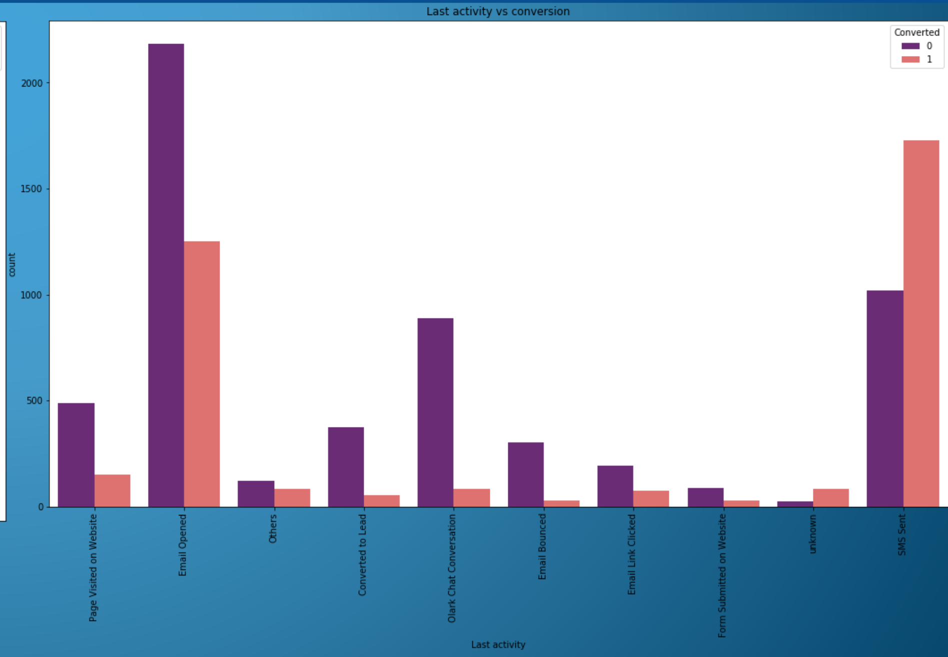
Tags:

1. many customers tend to get back after going through the email.
2. Closed by horizon has high lead conversion rate.

Last activity:

1. leads with last activity as SMS sent tend to have high conversion rate.
2. more number of leads tend to have their emails opened as the last activity.





Outlier Treatment

Points to be concluded from the graph.

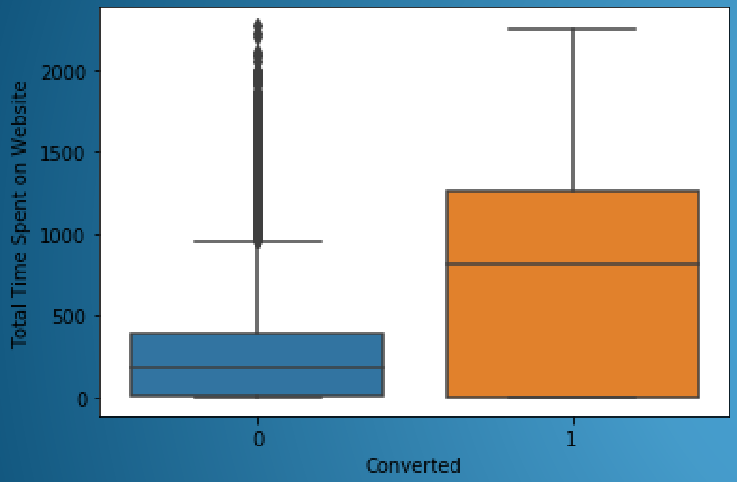
Total visits and page views per visit:

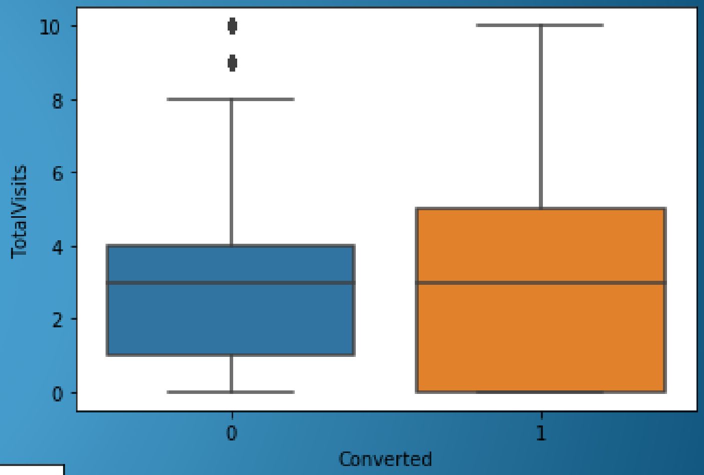
1. soft capping the outliers to 95% of the values
2. Both the medians are on the same level and nothing much can be inferred from above.

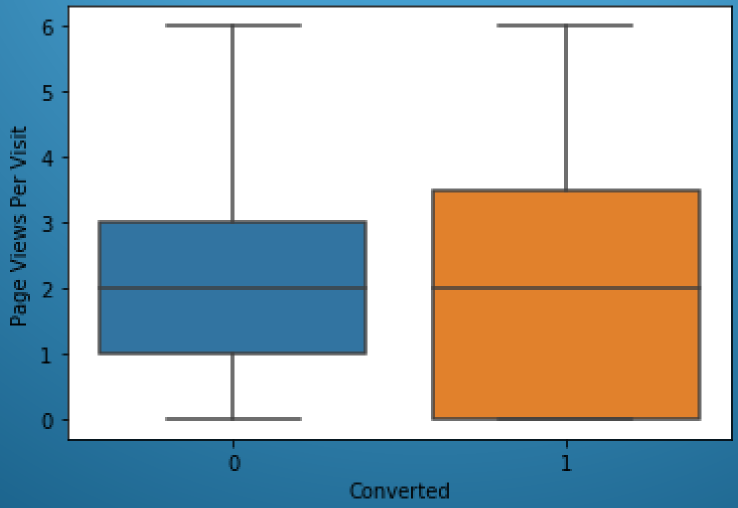
Total time spent on website

1. customers spending more time on the website tend to convert more.

2.Hence the customer experience on the website should be made more seamless and attractive.





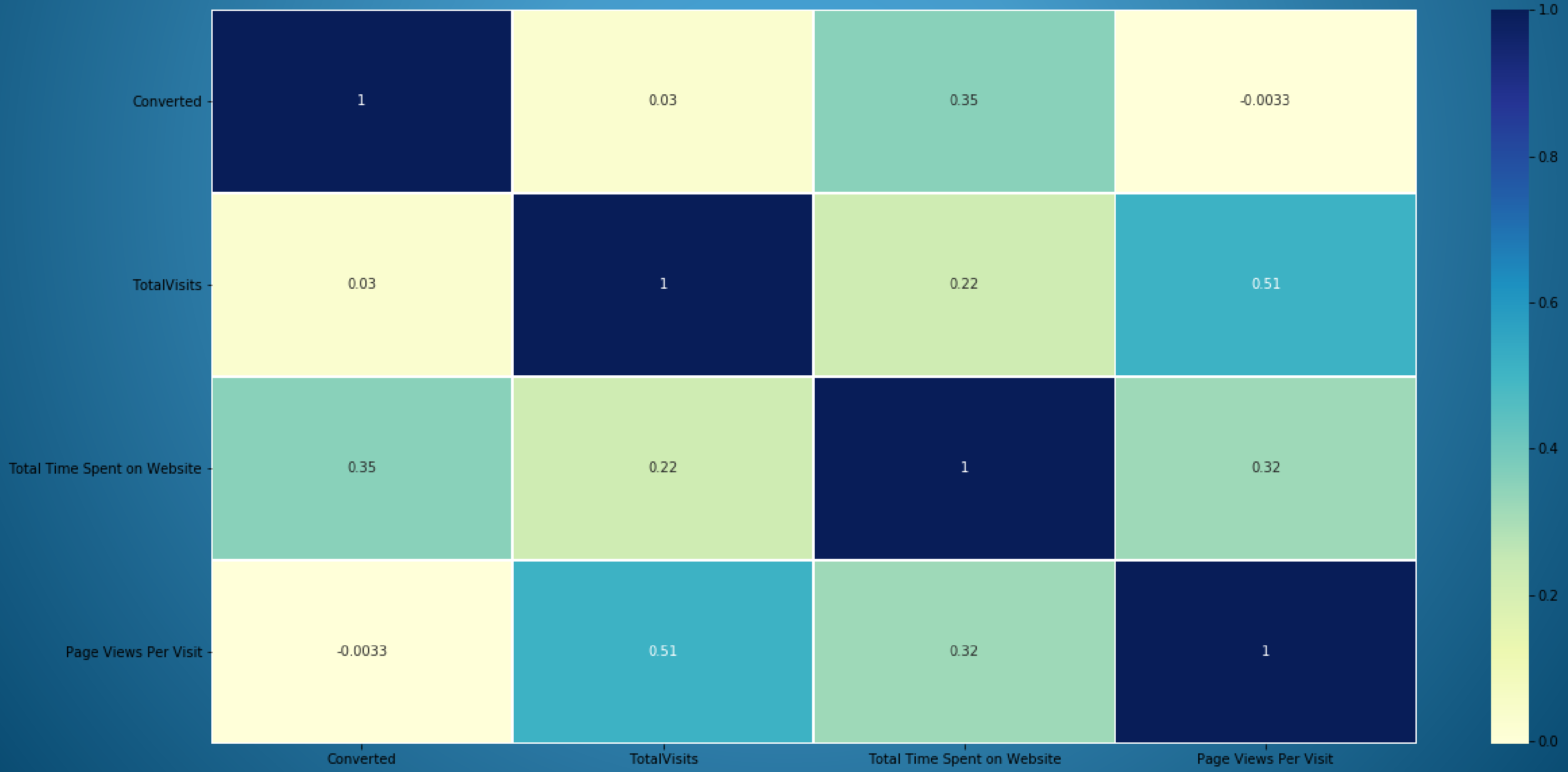


Numeric variable analysis

Points to be concluded from the graph.

1.Total visits and page views per visit are highly correlated.

2.Time spent on website has a 35% rate of customers getting converted into lead.



Data Preparation

Points to be noted.

1. Binary mapping of categorical variables.
2. Dummy Variable creation.
3. Dropping repetitive columns.
4. Total number of Rows : 9103 and Total number of columns : 60

Data Modeling

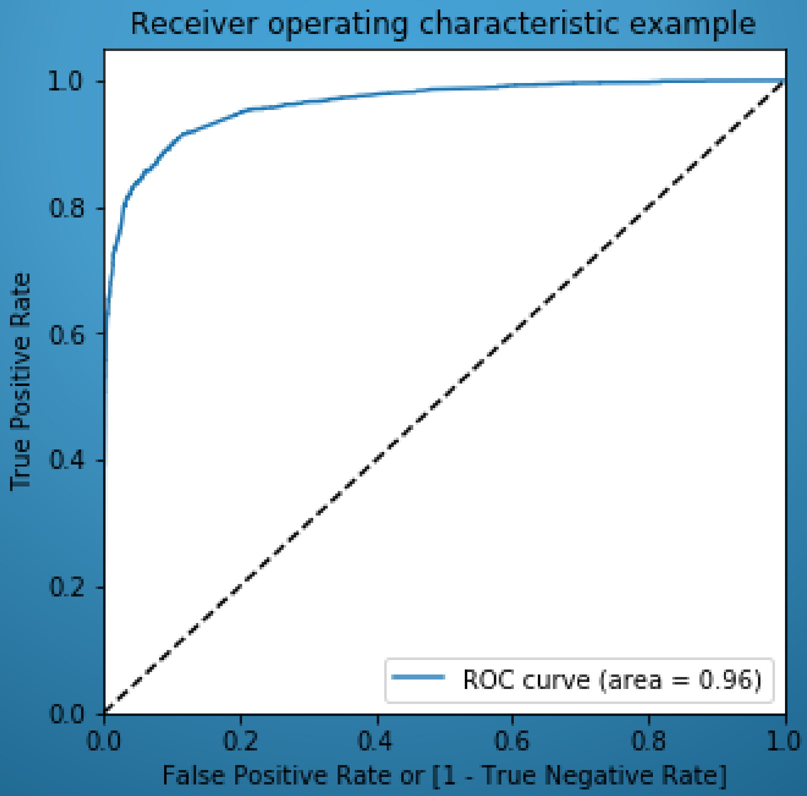
Points to be noted.

1. Splitting the Data into Training and Testing Sets: 70:30 ratio.
2. Use RFE for Feature Selection.
3. Running RFE with 15 variables.
4. Building Model by removing the variable whose p-value is greater than 0.05 and vif value is greater than 5.

ROC Curve

Points to be concluded from the graph.

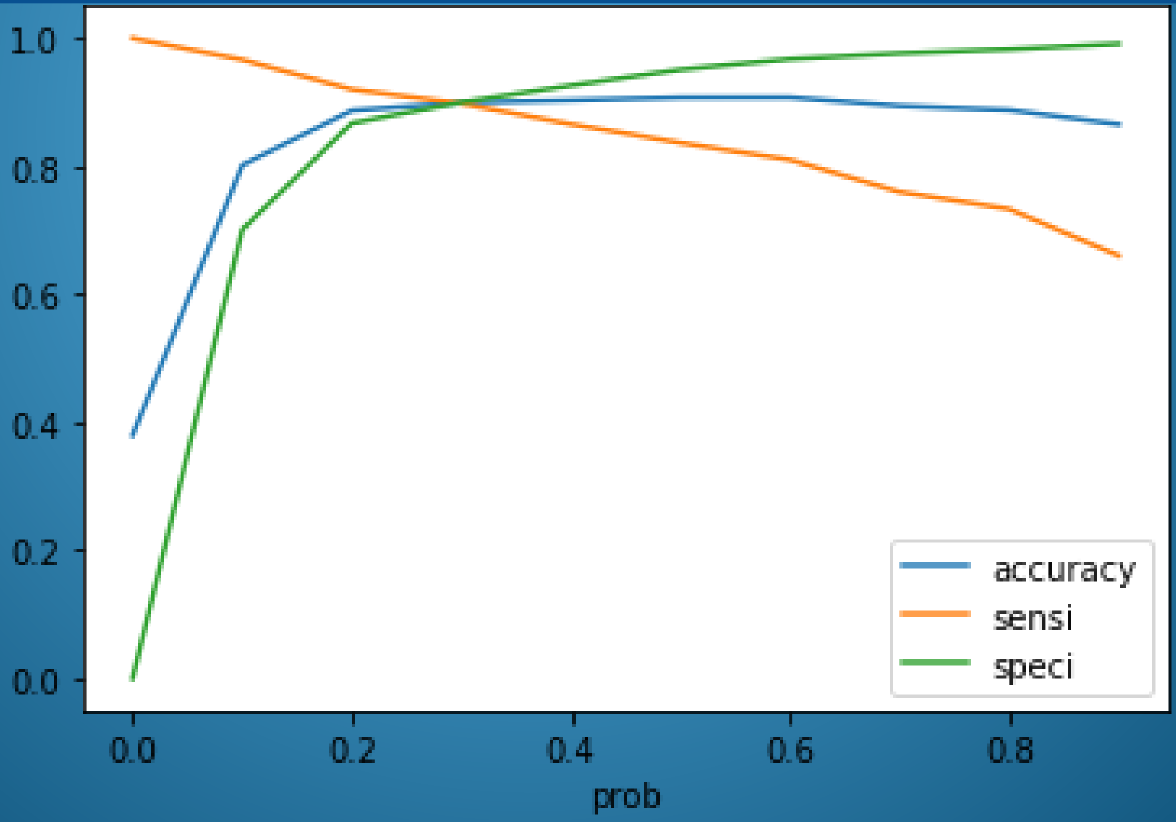
1. Calculate confusion matrix.
2. Calculate metrics like sensitivity, specificity, precision, Recall.
3. ROC curve demonstrates
4. It shows the tradeoff between sensitivity and specificity
5. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
6. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Optimal cut-off value

Points to be concluded from the graph.

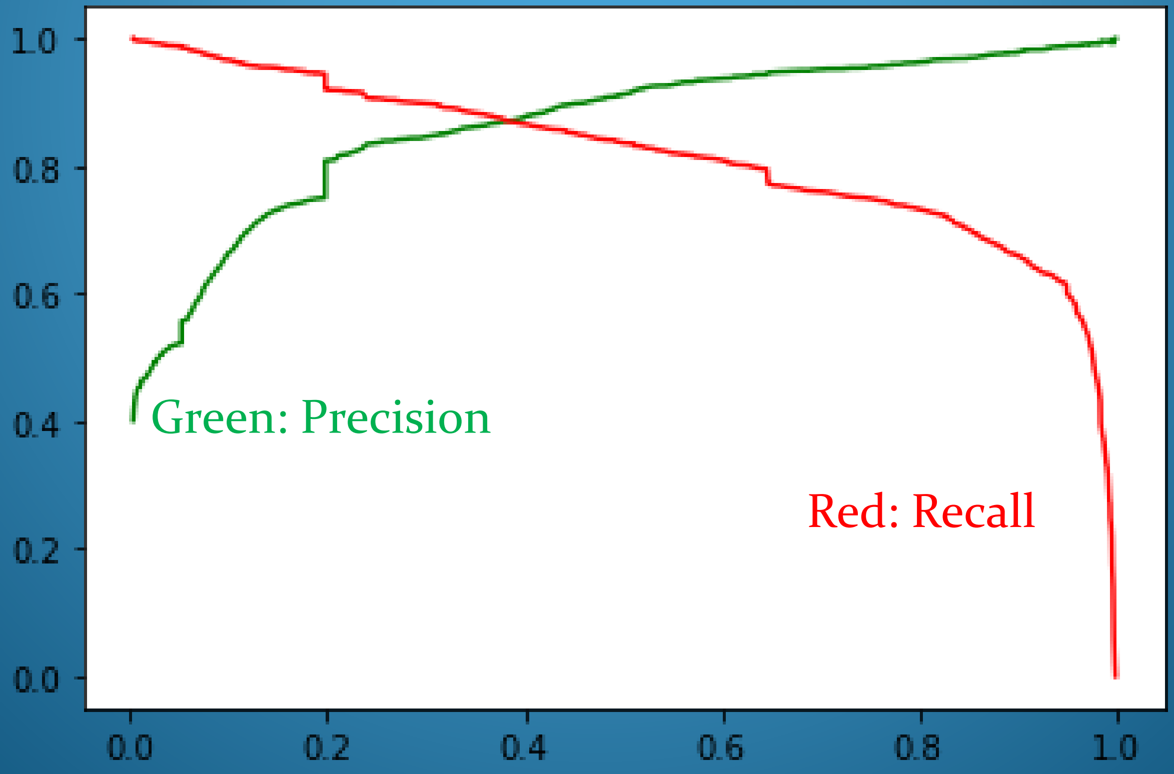
1. Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
2. From the second graph it is visible that the optimal cut off is at 0.30.



Precision Recall trade-off

Points to be concluded from the graph.

1. Cut-off value from the trade off seems to be at 0.38
2. . This value will be used to make predictions on the Test Set.



Comparison between Train and Test set.

Points to be concluded from the graph.

1. Insights: Train set data : with 0.3 cut –off

   Accuracy : 89%

   Sensitivity: 89%

   Specificity: 89%

   Precision :84%

   Recall : 89%

2. Insights: Test set data : with 0.38 cut –off

   Accuracy : 91%

   Sensitivity: 88%

   Specificity: 93%

   Precision :89%

   Recall : 88%

Conclusion

Points to be concluded from the Case Study

1. The model we choose has low p-values<0.05.
2. The model has very low VIF's <5 meaning very less multicollinearity among the variables.
3. The overall results on the Test sets seems to be good with accuracy of 91% , Sensitivity: 88% ,Specificity: 93% ,precision :89% ,Recall : 88%
4. Based on the model we have shortlisted the variables that contribute towards potential lead conversion rate.
5. High positive coefficients:

Tags Closed by Horizon.
Tags Will revert after reading the email.
Tags Busy.

Based on insights from the EDA analysis:

1. customers who are Unemployed are the ones who enquire more about X education online.
2. working professionals tend to get converted more.
3. google and direct traffic seems to have high number of leads.
4. wellingak website and reference tends to have a high conversion rate.
5. leads with last activity as SMS sent tend to have high conversion rate.
6. more number of leads tend to have their emails opened as the last activity.
7. API and Landing page submission tend to have higher numbers and better lead conversion rate.
8. Lead Add form has a higher lead conversion rate but lesser in numbers.
9. customers spending more time on the website tend to convert more.

Thank You