Lead Scoring Case Study

Submitted By:-

Vinay Patil

Business Understanding

1. 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.

2. The company markets its courses on several websites and search engines. The people land on the website and up a form providing their details, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

Problem statement

1. 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%.

2. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. 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.

Business Objectives

1. The company requires to build a model to get most potential leads that are most likely to convert into paying customers.

2. Assign a lead score (between 0 and 100) to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

3. The lead conversion rate of the model souled be greater than 80 % i.e. company wants at build the model with 80 % accuracy

Analysis Approach

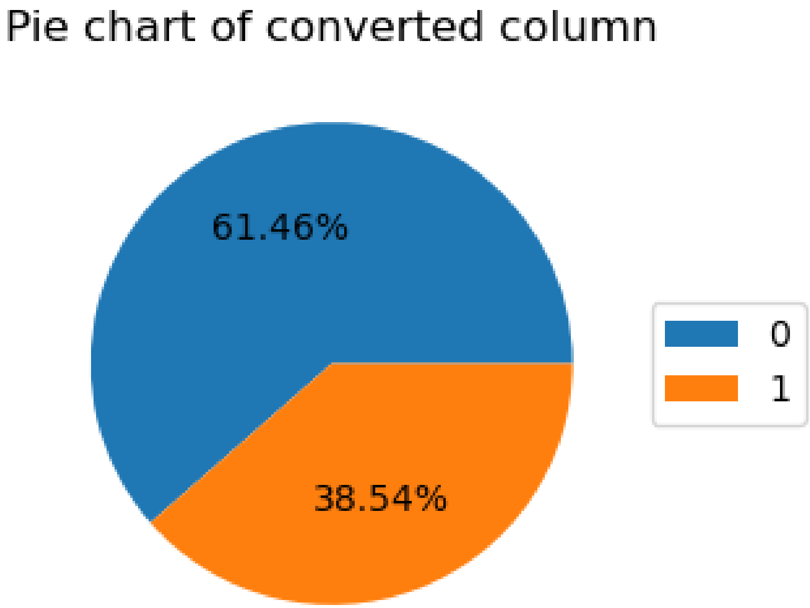
1. Import the data for analysis.
2. Understand the data i.e. understand each column,
3. Cleaning and preparing the data for model building:
    1. Dropping duplicated data
    2. Delete column with higher missing value
    3. Imputing missing value with appropriate mean or median.
    4. Deleting non relative column i.e. column having all unique value (categorical column) or only one unique value
    5. Handling outliers and imputing threshold value if require
4. Perform Exploratory Data Analysis.
    1. Univariate data analysis
    2. Bivariate data analysis
5. Perform feature scaling and impute dummy variable
6. Splitting the data into Test and Train dataset.
7. Building a logistic Regression model and calculate Lead Score.
8. Finding optimum cutoff lead score
9. Get Confusion matrix and calculate accuracy, Sensitivity and Specificity, Precision and Recall
10. Apply the model on test data and validate based on Confusion matrix and calculate accuracy, Sensitivity and Specificity, Precision and Recall

Data Cleaning and preparing

1. Remove the duplicate data.
2. Making lead number as index to retain the identity of the leads
3. Removing Prospect ID
4. Imputing select value in the data with null or not specified
5. Removing column with missing value percentage greater than 40%
6. Removing biased column which is having 95% biased value, as this will leads to incorrect decision
7. Imputing outliers with threshold value = Q3+1.5*IQR
   - Q3 = 75% Quantile
   - IQR = inter Quantile range
8. Merging the values with low frequency as Others

Exploratory Data Analysis

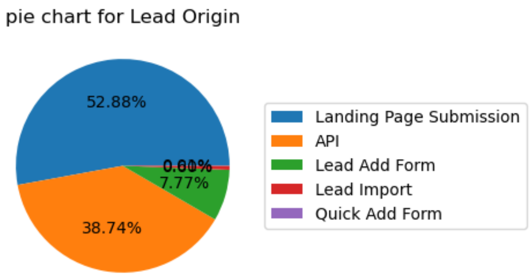
Target Column

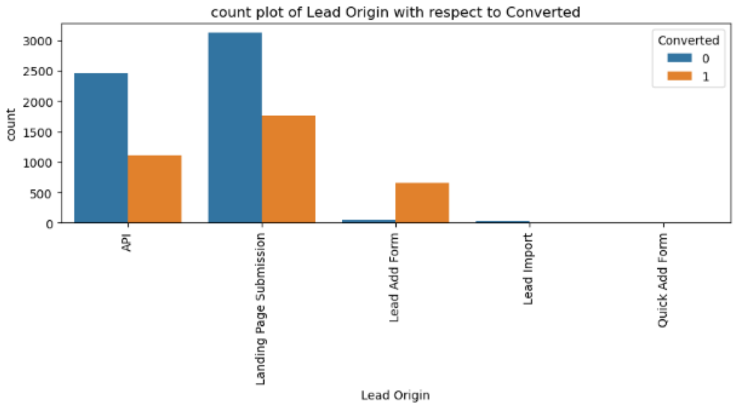


Inference

1. Data is balanced
2. We are having 61% leads who are not converted

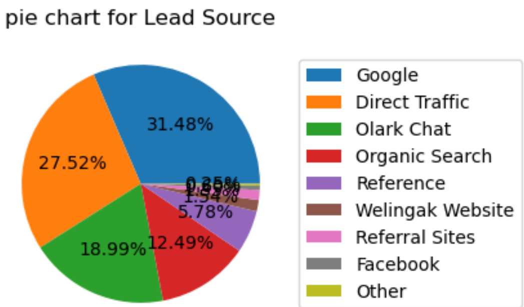
Exploratory Data Analysis

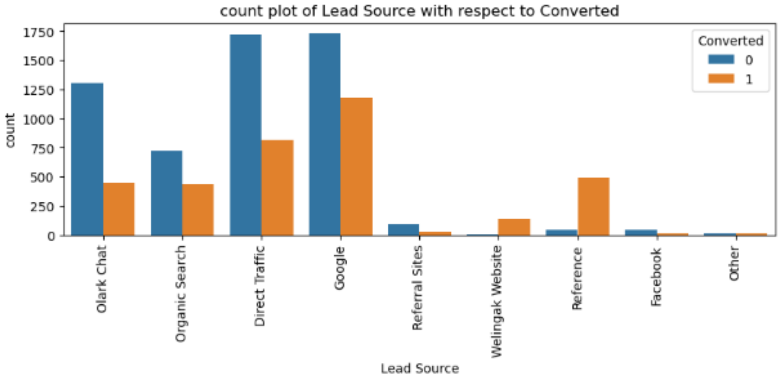




Inference

1. 'API' and 'Landing Page Submission' has higher number of leads as well as 'conversion’.
2. 'Lead Add Form' has a very high conversion rate but count of leads are not very high.
3. To increase the conversion we have to increase the 'Lead Add Form'

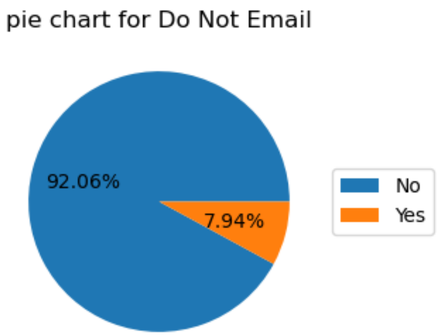


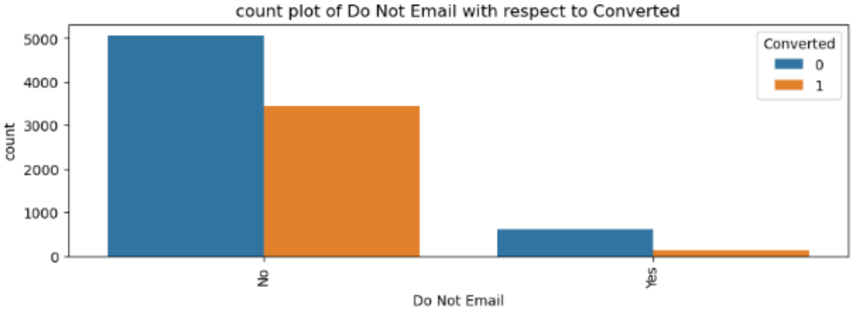


Inference

1. 'Olark Chat', 'Organic Search', 'Google', 'Direct Traffic' has higher number of leads as well as 'conversion’.
2. 'Reference' has a very high conversion rate.
3. To increase the conversion we have to increase the 'Reference'

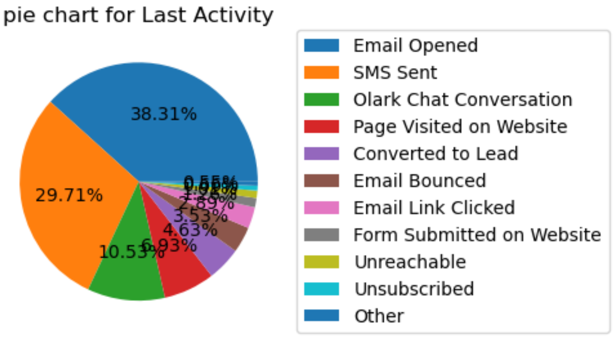
Exploratory Data Analysis

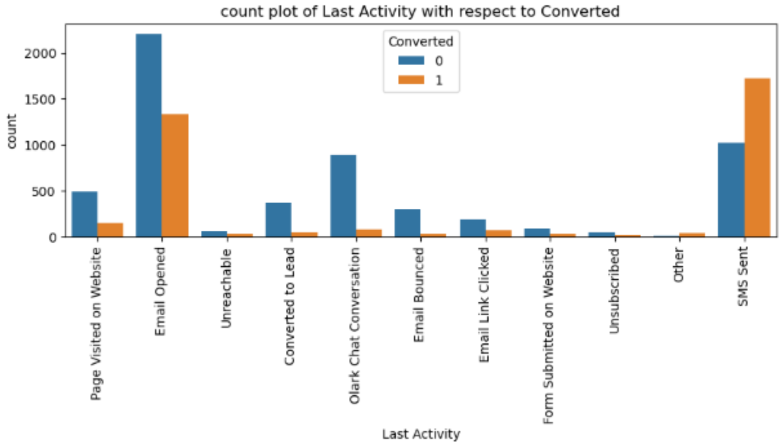




Inference

1. Data us biased here, however we can delete column which are having more than 95% biased data
2. Lead Who opted 'no' to the 'do not email' had more number of leads

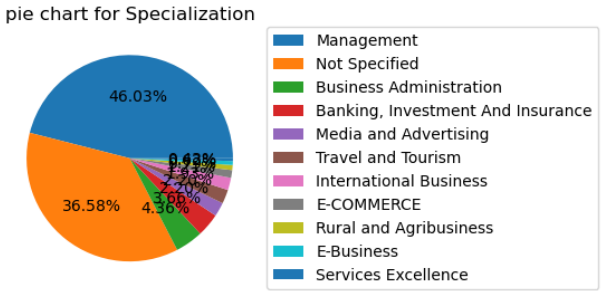


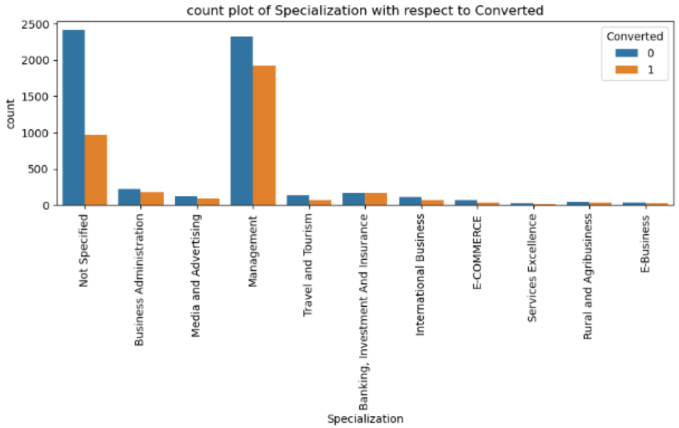


Inference

1. Data is not biased
2. "email opened" and “SMS sent" has more number of leads
3. “SMS sent" has high rate of lead conversion

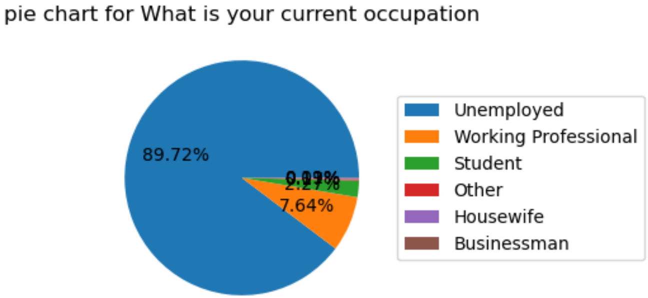
Exploratory Data Analysis

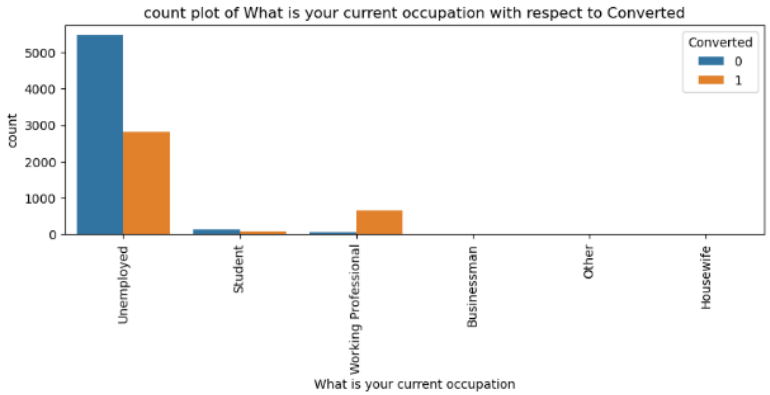




Inference

1. Management specialization have higher number od leads and conversion rate
2. Banking, Investment And Insurance has almost 100% conversion so we can focused more on management and Banking, Investment And Insurance

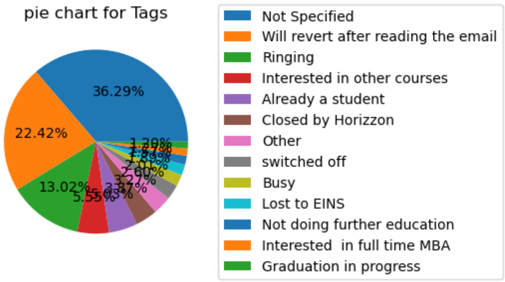


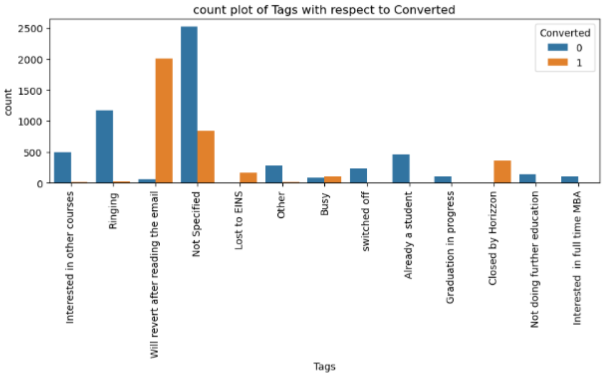


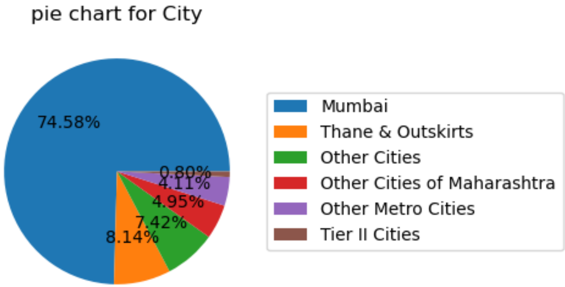
Inference

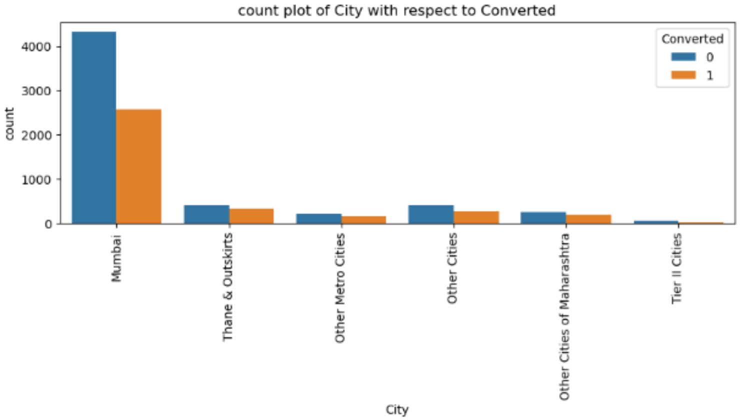
1. Unemployed has higher number of leads with approximate 50% conversion ratio
2. Working Professionals have high chances of conversion

Exploratory Data Analysis









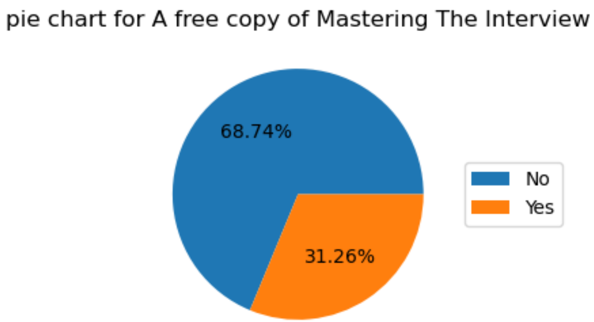
Inference

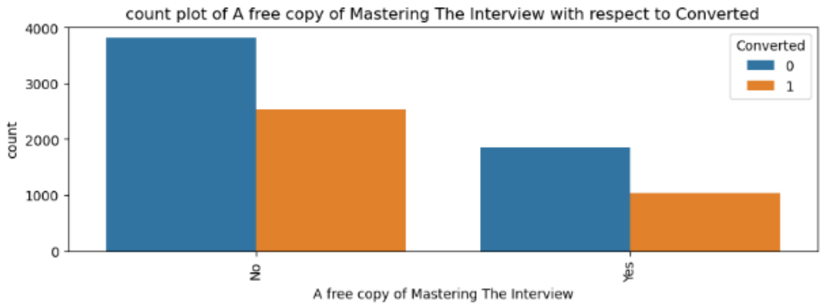
1. 'Will revert after reading the email' will generate more number of leads and higher conversion
2. "Closed by Horizzon" has higher conversion rate

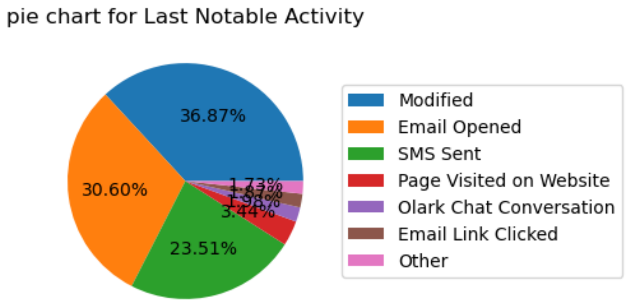
Inference

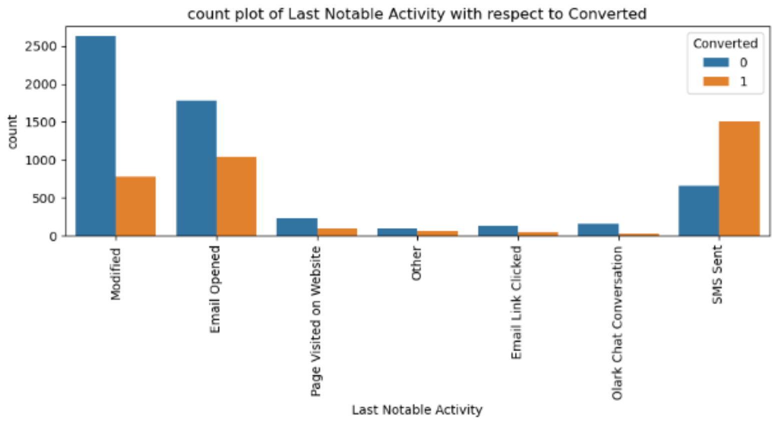
1. higher number of leads are generated from Mumbai
2. conversion rate is almost same in all the city

Exploratory Data Analysis









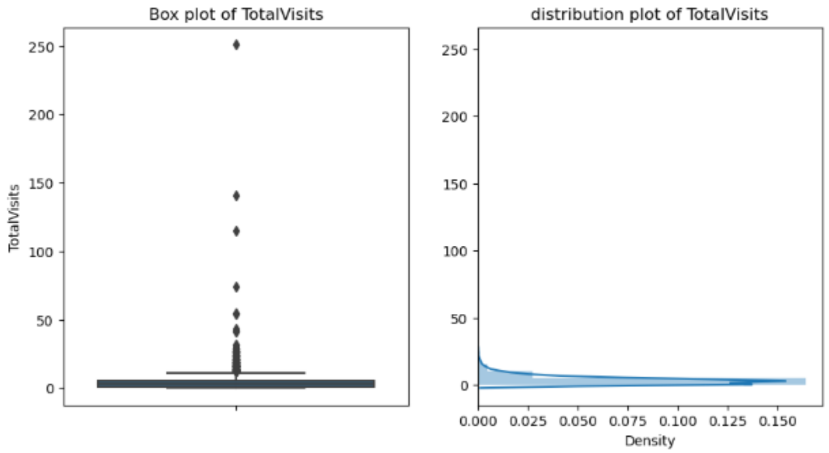
Inference

1. higher number of leads are generated who answered 'no' to 'A free copy of Mastering The Interview’
2. conversion rate is almost similar

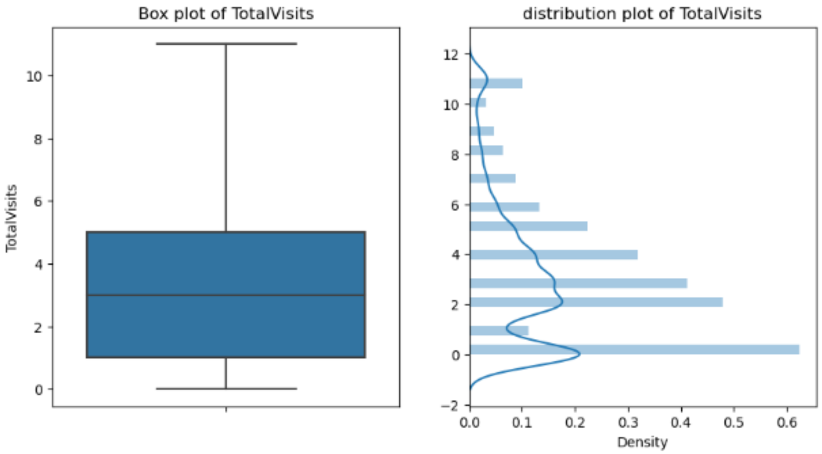
Inference

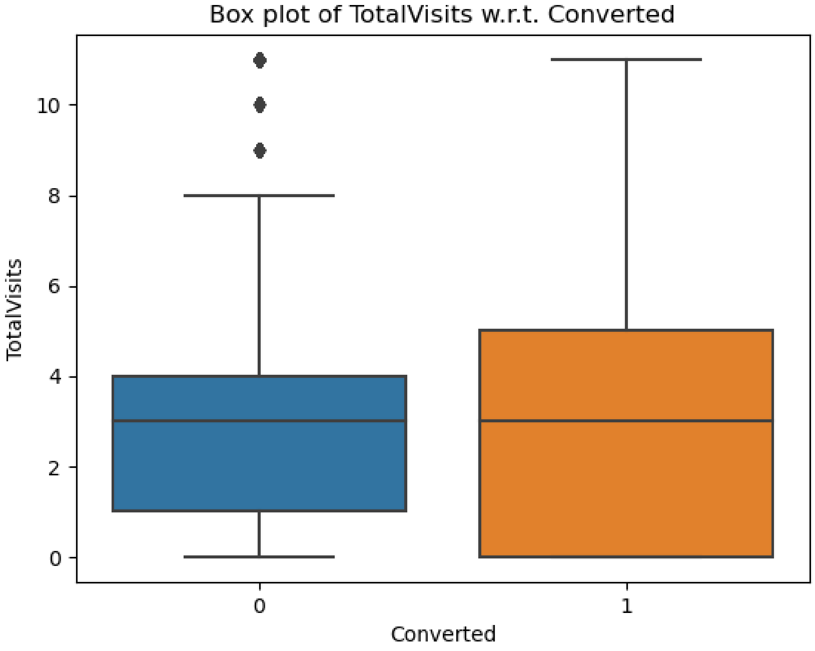
1. 'Modified', 'Email Opened ','SMS Sent' generate higher number of leads
2. SMS sent are having heigher conversion rates

Exploratory Data Analysis for TotalVisits



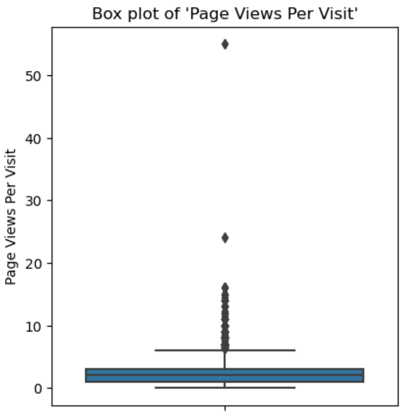
Imputed 267 outliers with threshold value of the data i.e. 11

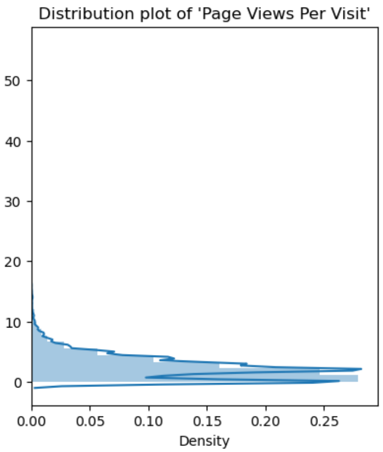




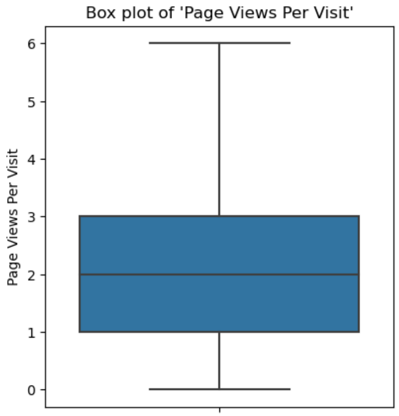
1. Median for converted and not converted leads are the same

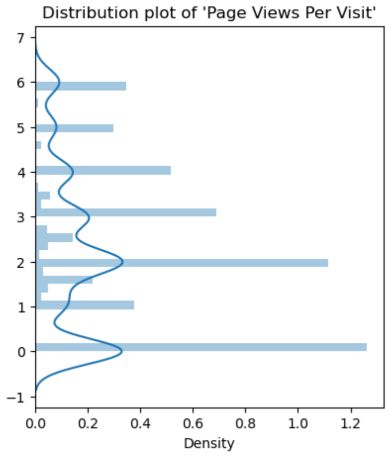
Exploratory Data Analysis for “Page Views Per Visit”

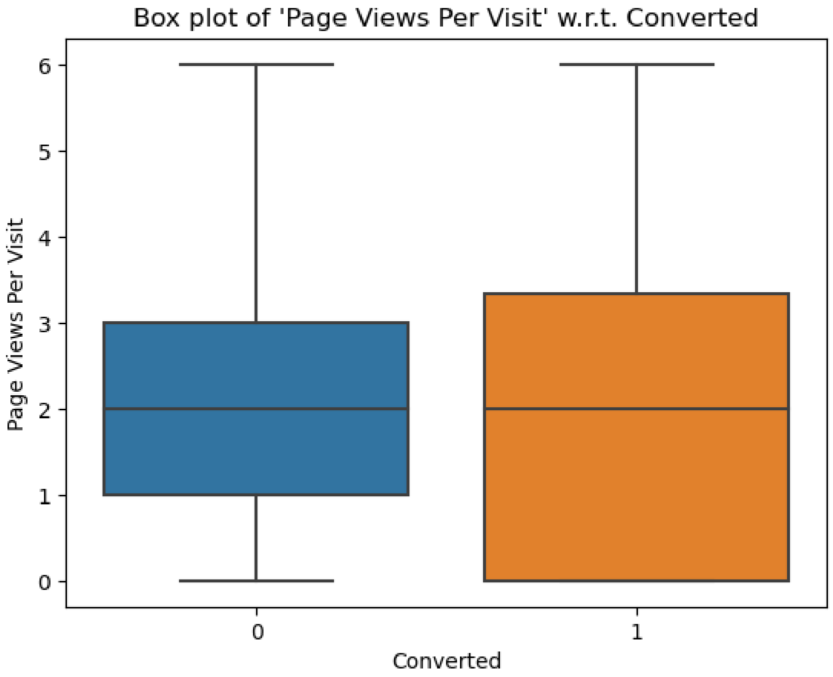




Imputed 360 outliers with threshold value of the data i.e. 6



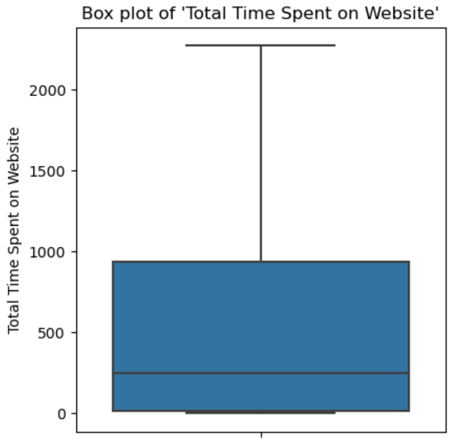


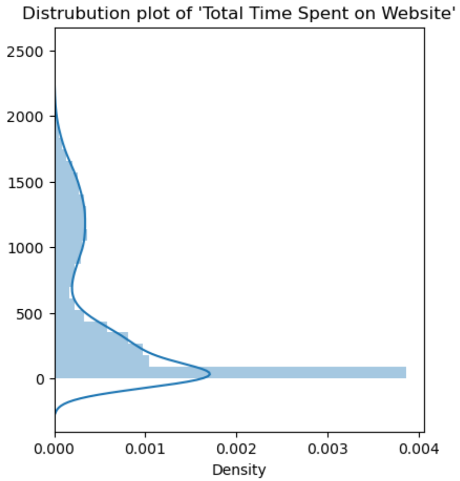


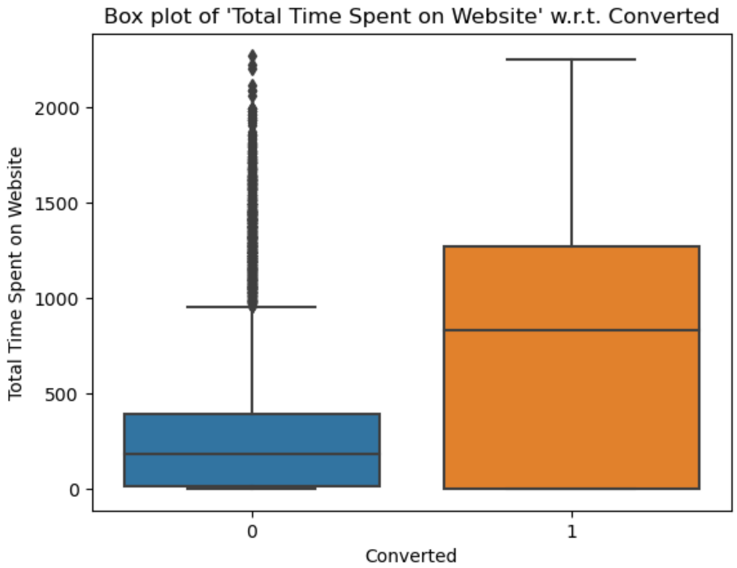
Inference

1.Median for converted and not converted leads are the same

Exploratory Data Analysis for “Total Time Spent on Website ”


Box plot of 'Total Time Spent on Website'

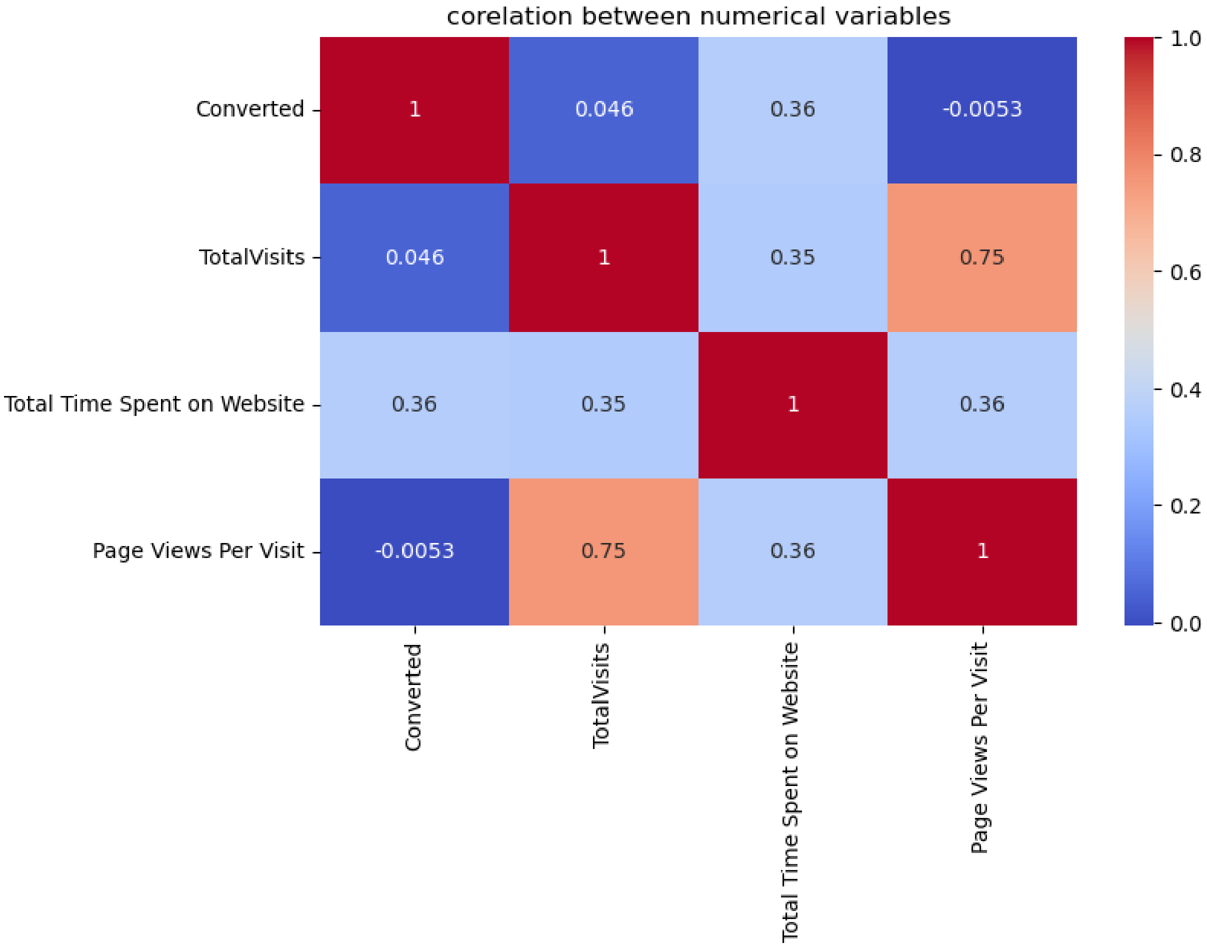

Distrubution plot of 'Total Time Spent on Website'


Box plot of 'Total Time Spent on Website' w.r.t. Converted

Inference

1.leads who are spending more time will mostly to be converted.'

Exploratory Data Analysis



Data Preparation for Model Building

1. After creating dummy variable, we are having 9240 rows and 66 columns
2. Splitting the data into train and test set with 70:30 ration
3. Scaling the train data using Standard Scaler

Model Building

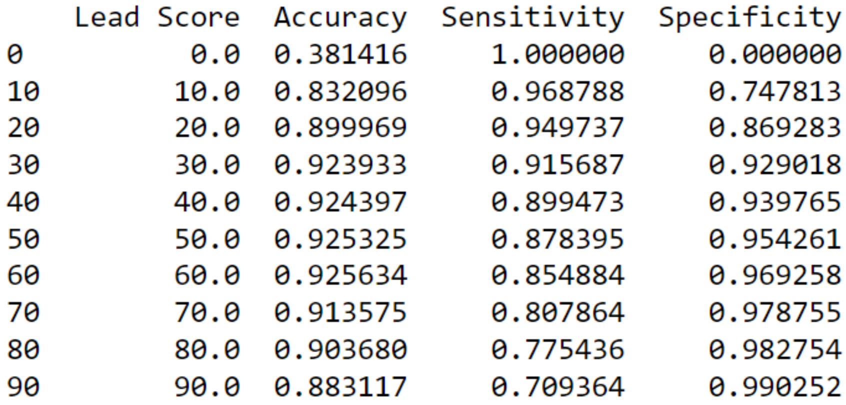
1. Recursive feature elimination (RFE) and retain 20 columns

2. Building Logistic regression model and removing column with p-value > 0.05 and VIF > 5, we have final model with 19 columns.

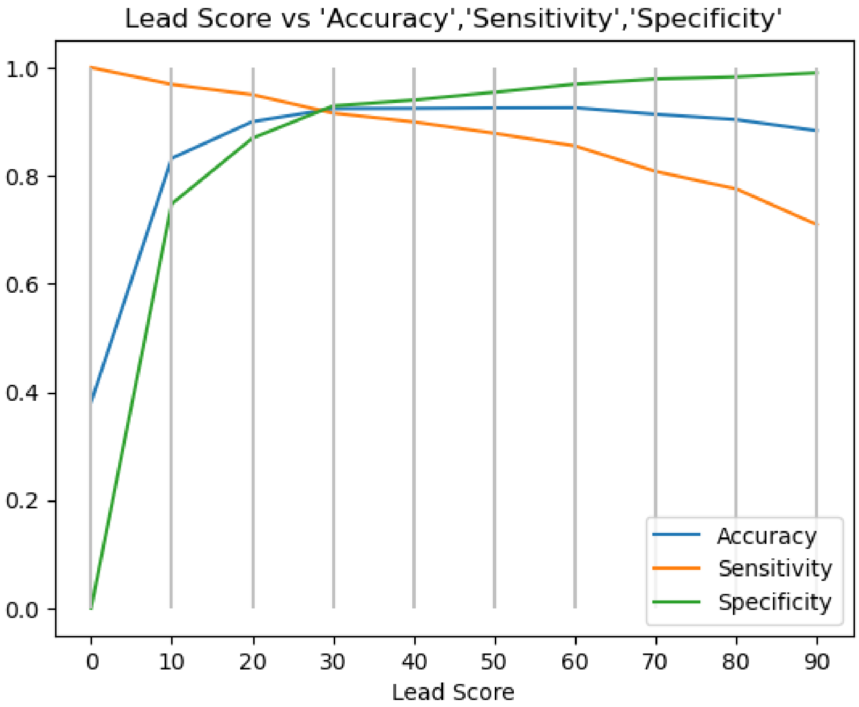
   ['Do Not Email', 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'Tags_Already a student', 'Tags_Closed by Horizzon', 'Tags_Interested in full time MBA', 'Tags_Interested in other courses', 'Tags_Lost to EINS', 'Tags_Not doing further education', 'Tags_Other', 'Tags_Ringing', 'Tags_Will revert after reading the email', 'Tags_switched off', 'Last Notable Activity_Email Link Clicked', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversation'],

Model Evaluation Train Data set

Finding Optimal Cutoff Point



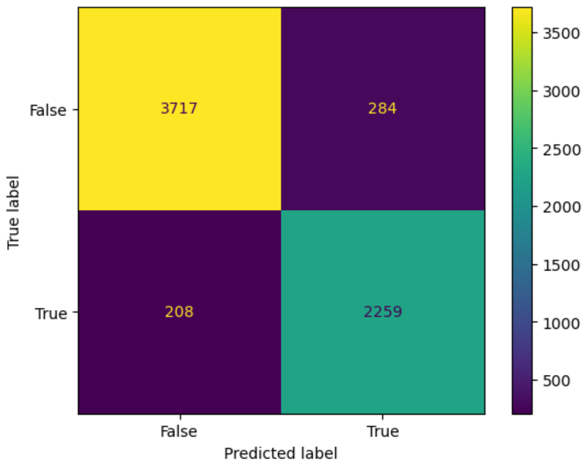
```
    Lead Score  Accuracy  Sensitivity  Specificity
0          0.0  0.381416     1.000000     0.000000
10        10.0  0.832096     0.968788     0.747813
20        20.0  0.899969     0.949737     0.869283
30        30.0  0.923933     0.915687     0.929018
40        40.0  0.924397     0.899473     0.939765
50        50.0  0.925325     0.878395     0.954261
60        60.0  0.925634     0.854884     0.969258
70        70.0  0.913575     0.807864     0.978755
80        80.0  0.903680     0.775436     0.982754
90        90.0  0.883117     0.709364     0.990252
```

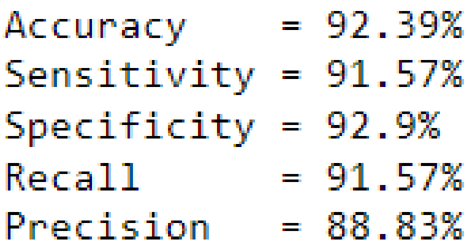


From the curve above, 30 is the optimum point to take as a cutoff of lead score.

Model Evaluation Train Data set

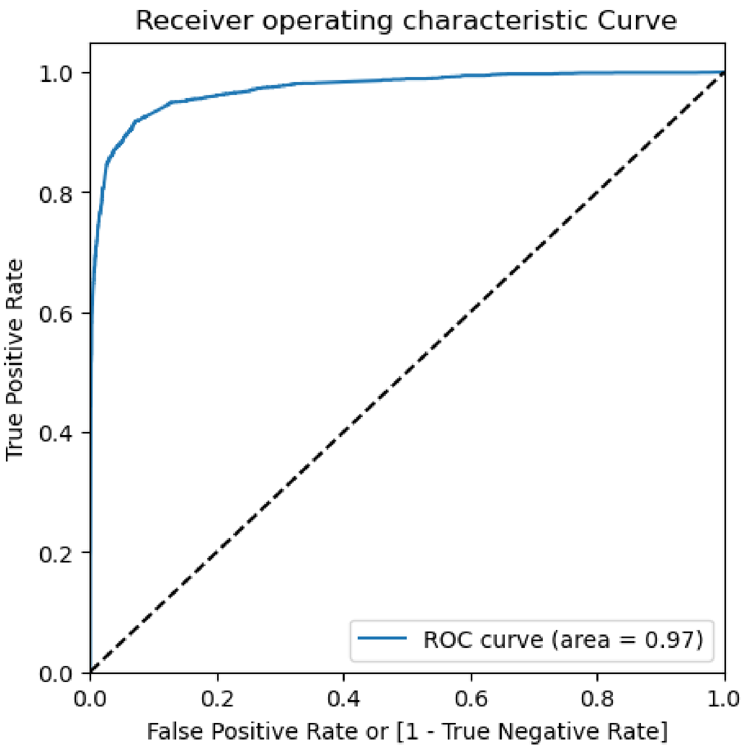
Confusion Matrix





```
Accuracy    = 92.39%
Sensitivity = 91.57%
Specificity = 92.9%
Recall      = 91.57%
Precision   = 88.83%
```

ROC Curve

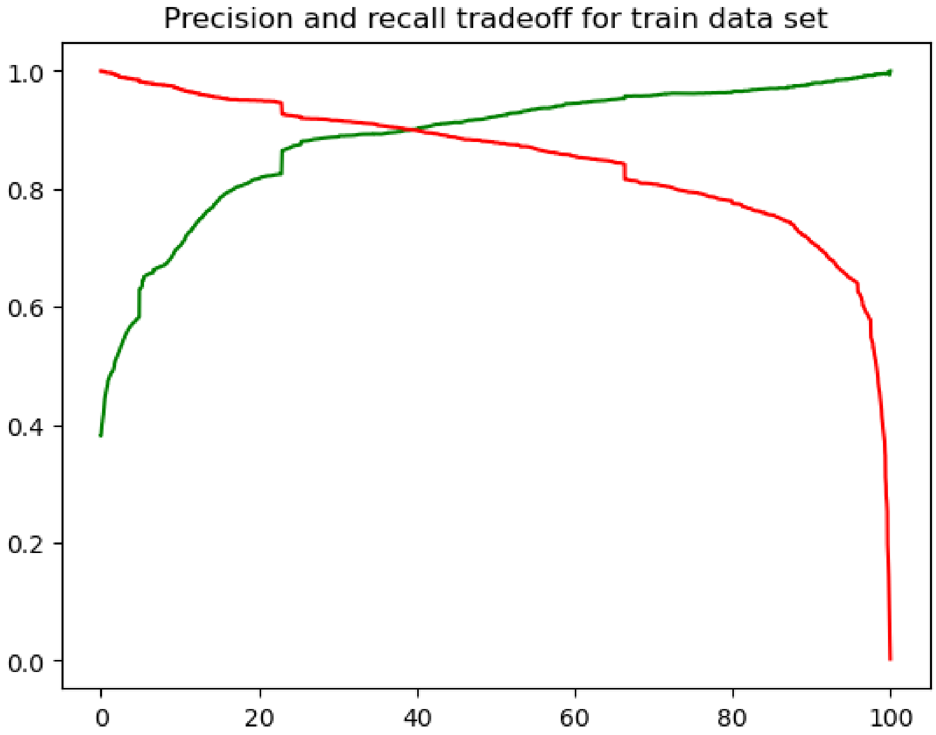


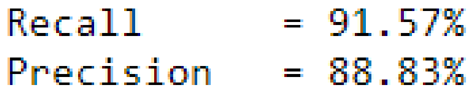
Inference

1. The area Under ROC curve should be close to 1 for 100% predictive model
2. We are getting area Under ROC curve = 0.97 which is close to 1 this indicating a good predictive model

Model Evaluation Train Data set

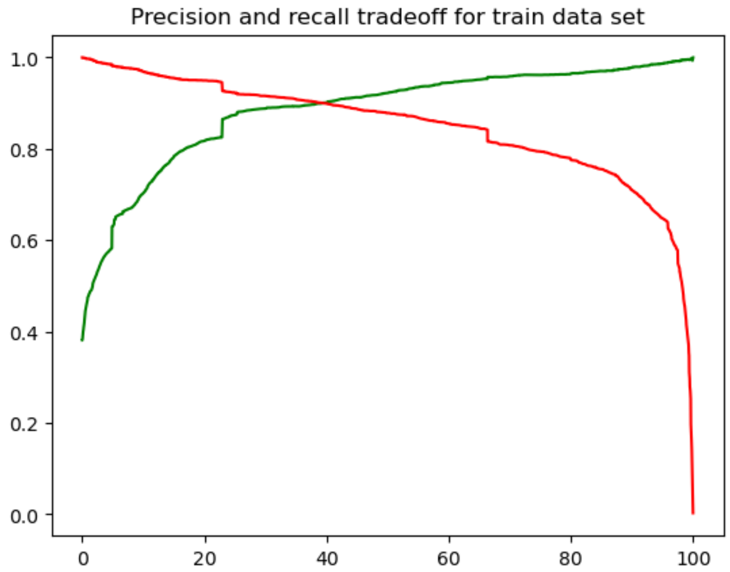
Precision and recall tradeoff for train data set

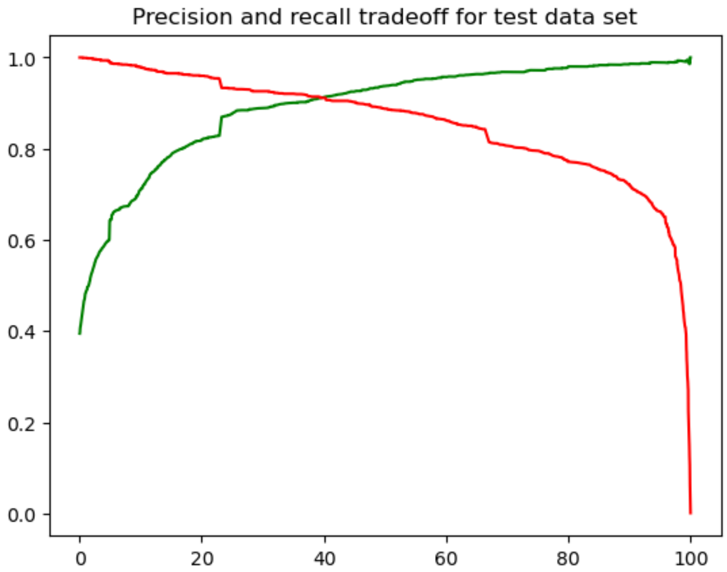




```
Recall       = 91.57%
Precision    = 88.83%
```

Model Evaluation Comparison between test and train data





	Train Data Prediction	Test Data Prediction
Accuracy	92.39	92.53
Sensitivity	91.57	92.60
specificity	92.90	92.49
Recall	91.57	92.60
Precision	88.83	88.94

Comments

1. we have achieved our goal of the accuracy >80%
2. Based of the above model our lead conversion rate is 92% accurate

Conclusion

1. We have considered the optimal cut off of lead score based on Accuracy, Sensitivity and Specificity for calculating the final prediction.

2. We have also visualize the trade off between precision and recall.

3. The final prediction of Conversion rate for train data is 92.39 % and for test data is 92.53 %.

4. The variables which contribute most towards the probability of a lead getting converted is:
   1. Lead Source
   2. Last Activity
   3. Lead Origin

5. Over all Accuracy of the model is more than 80% hence we have achieved our Goal