LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

PROBLEM STATEMENT

- 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%.

BUSINESS GOALS

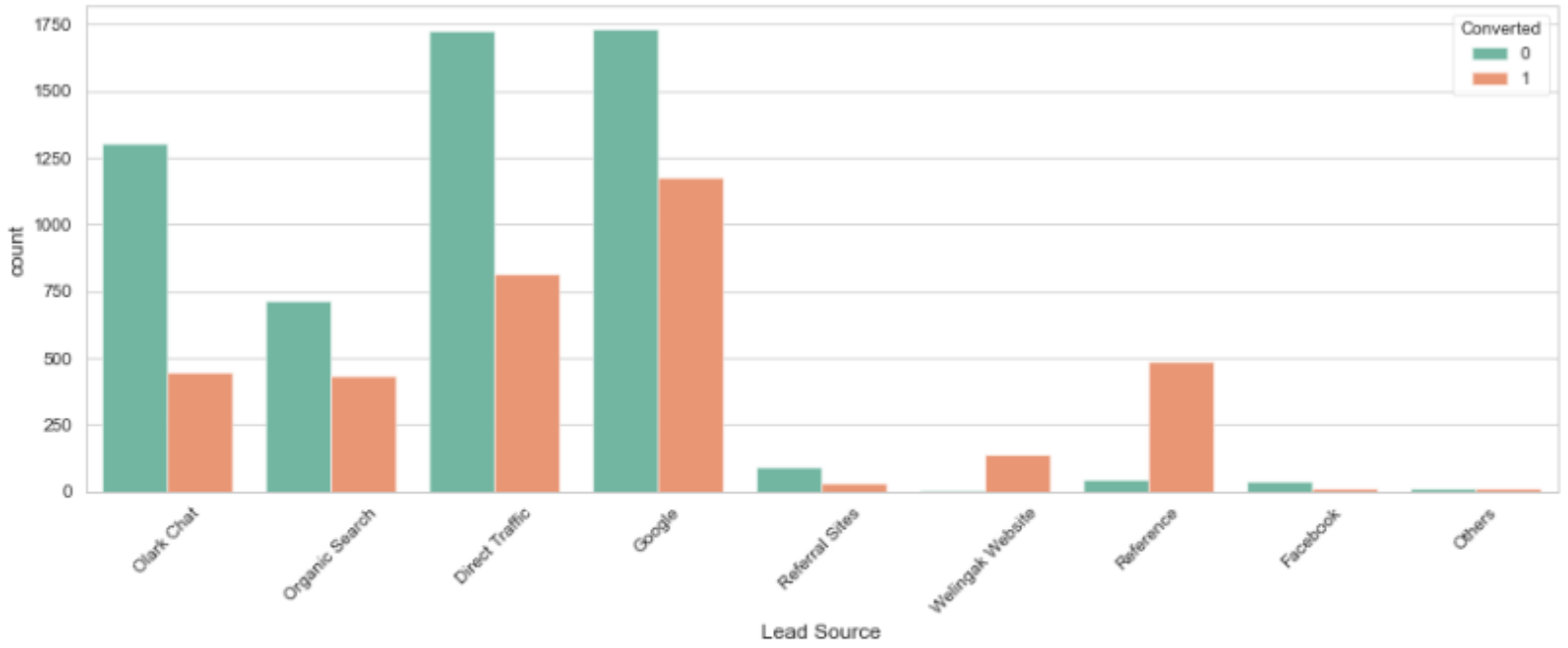
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor.
- For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- 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.

ANALYSIS APPROACH

- 1. Reading and Understanding the data
- 2. Inspecting Dataframe
- 3. Missing Value Treatment
- 4. Data Preparation
- 5. EDA
- 6. Convert Binary Categories
- 7. Dummy Variable
- 8. Train- Test Split
- 9. Model Building
- 10. Model Evaluation : Train Dataset
- 11. Model Evaluation : Test Dataset
- 12. Conclusion

EXPLORATORY DATA ANALYSIS

- **LEAD SOURCE**



Inference

Maximum Leads are generated by Google and Direct Traffic.

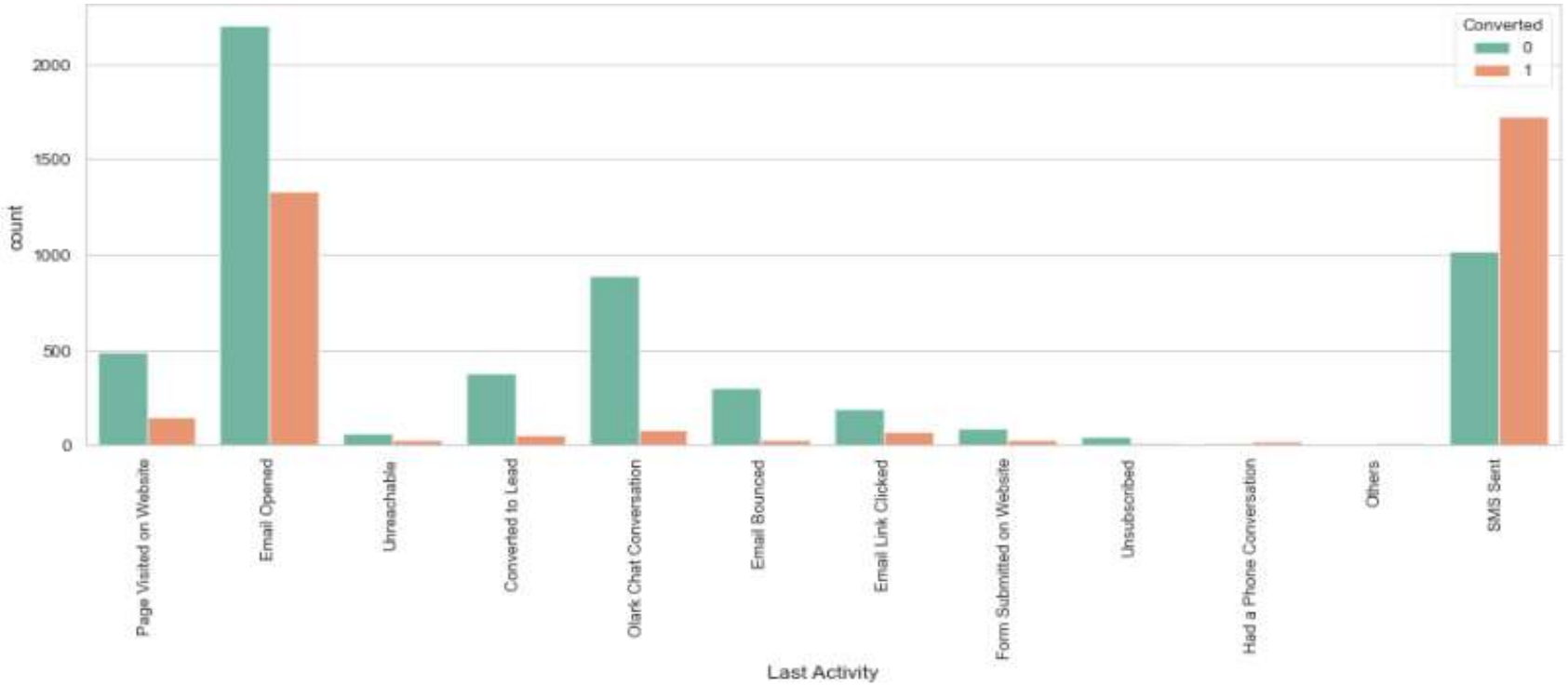
Conversion rate of Reference leads and Welinkgak Website leads is very high.

ANALYSIS

- The company should focus on getting more references as the conversion rate is very high.
- Also, the leads coming from Welinkgak Website has a high conversion rate, therefore should promote more on that website.
- Although the most number of leads are generated from Google and Direct traffic but the conversion rate is not good, therefore the team should focus more on nurturing these leads.

EXPLORATORY DATA ANALYSIS

- **LEAD ACTIVITY**



Inference

Maximum leads are generated having last activity as Email opened but conversion rate is not too good.

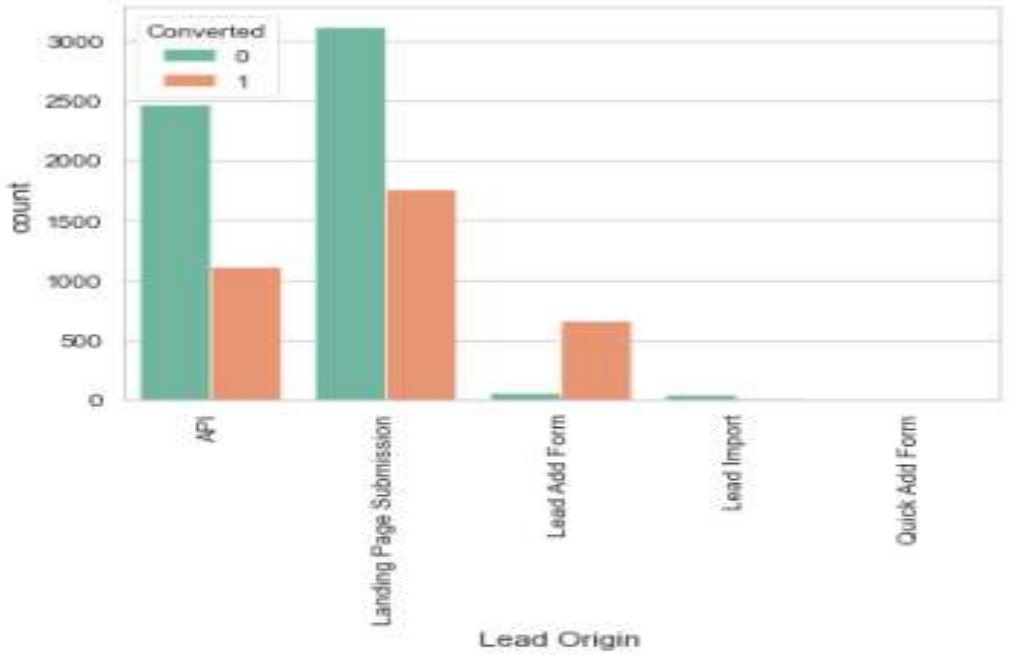
SMS sent as last acitivity has high conversion rate.

ANALYSIS

- The company should focus on sending more SMS as the conversion rate is very high.
- The most number of leads are generated having last activity as Email opened but the conversion rate is not good, therefore the team should focus more on conveying precise information through Emails so that their conversion rate increases.

EXPLORATORY DATA ANALYSIS

- **LEAD ORIGIN**



Inference

Maximum leads are generated through Landing Page Submission.

Lead Add Form has high conversion rate.

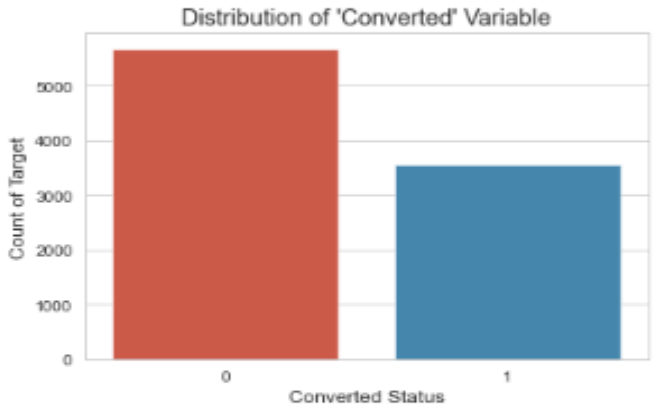
ANALYSIS

- The company should focus on promoting lead add form as the conversion rate is very high.

- The most number of leads are generated from Landing Page Submission but the conversion rate is not good, therefore the team should focus more on nurturing these lead so that their conversion rate increases.

NUMERIC ATTRIBUTE ANALYSIS

In [48]:
```
#Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0).
#Visualizing Distribution of 'Converted' Variable
sns.countplot(lead.Converted)
plt.xlabel("Converted Status")
plt.ylabel("Count of Target")
plt.title("Distribution of 'Converted' Variable")
plt.show()
```



In [49]:
```
# Finding out conversion rate
Converted = (sum(lead['Converted'])/len(lead['Converted'].index))*100
Converted
```

Out[49]: 38.53896103896104

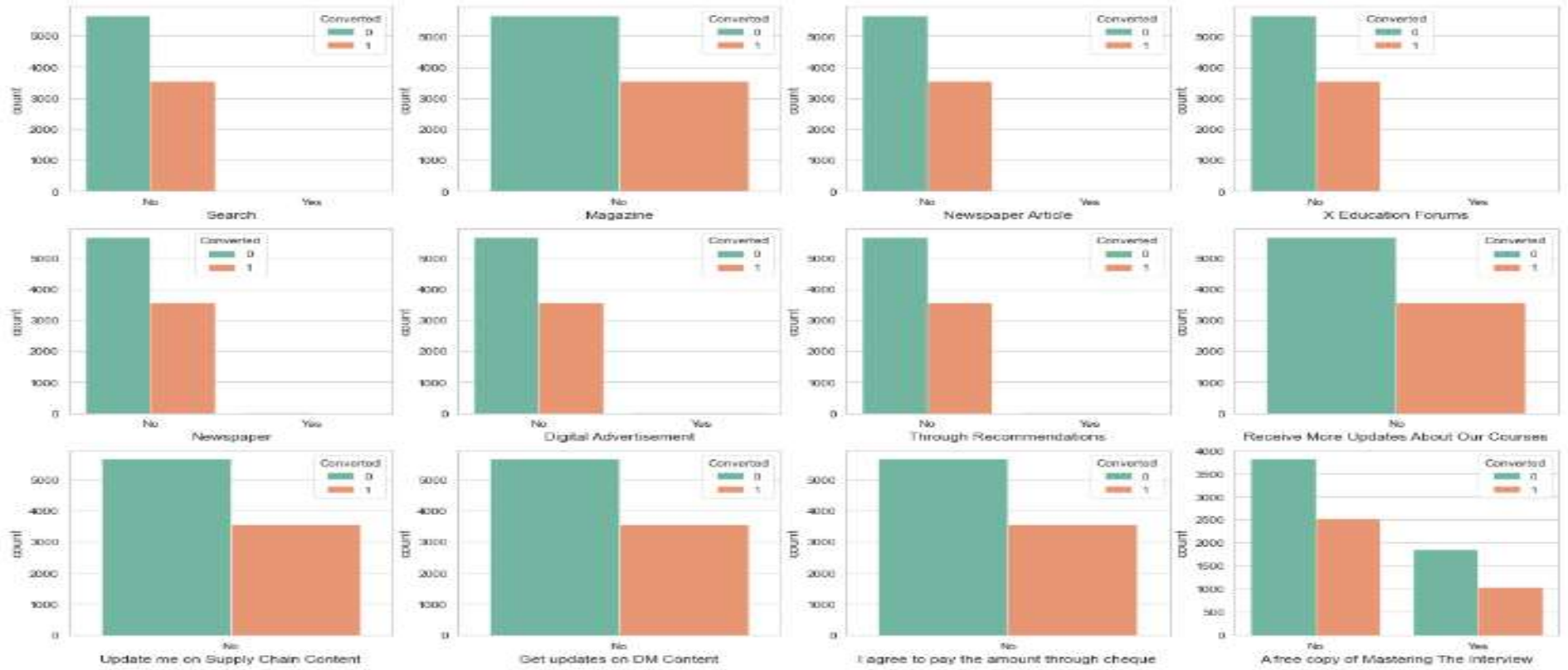
Inference

Currently, lead Conversion rate is 38% only

ANALYSIS

- The data set provided has the lead conversion rate of 38%
- Need to build a logistic regression model which gives the lead conversion rate as 80%

CATEGORICAL ATTRIBUTE ANALYSIS : IMBALANCED VARIABLES



Inference

For all these columns except 'A free copy of Mastering The Interview' data is highly imbalanced, thus we will drop them "A free copy of Mastering The Interview" is a redundant variable so we will include this also in list of dropping columns.

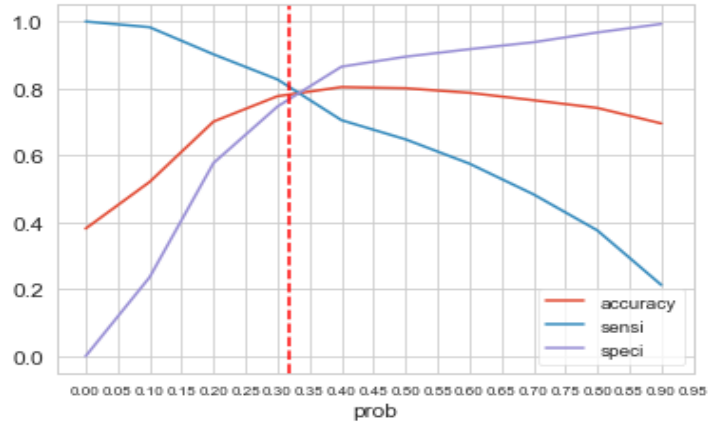
MODEL BUILDING

- SPLITTING INTO TRAIN AND TEST SET
- SCALE VARIABLE IN TRAIN DATA SET
- USE RFE TO ELIMINATE LESS RELEVANT VARIABLES
- BUILD THE NEXT MODEL
- ELIMINATE VARIABLES BASED ON HIGH p-VALUE
- CHECK VIF VALUE FOR ALL THE EXISISTING COLUMNS
- EVALUATE ACCURACY AND OTHER METRIC
- PREDICT USING TEST SET

Model Evaluation : Train Dataset

```
In [96]: # Let's plot accuracy sensitivity and specificity for various probabilities.
         plt.figure(figsize=(18,8))
         sns.set_style("whitegrid")
         cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
         plt.xticks(np.arange(0,1,step=0.05),size=8)
         plt.axvline(x=0.317, color='r', linestyle='--') # additing axline
         plt.yticks(size=12)
         plt.show()
```

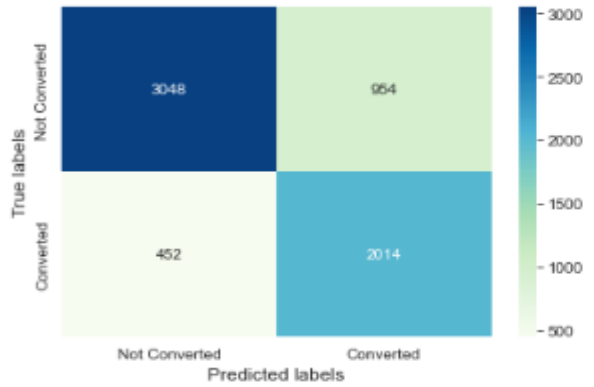
```
<Figure size 1296x576 with 0 Axes>
```



At the cut-off of 0.317 we can see the accuracy, sensitivity and specificity in the graph above

CONFUSION MATRIX

In [100]:
```
#Plotting the Confusion Matrix
draw_cm( y_train_pred_final['Converted_Value'], y_train_pred_final['final_predicted_1'], "GnBu")
```



In [101]:
```
conf_matrix = confusion_matrix(y_train_pred_final['Converted_Value'], y_train_pred_final['final_predicted_1'] )

lg_metrics(conf_matrix)
```

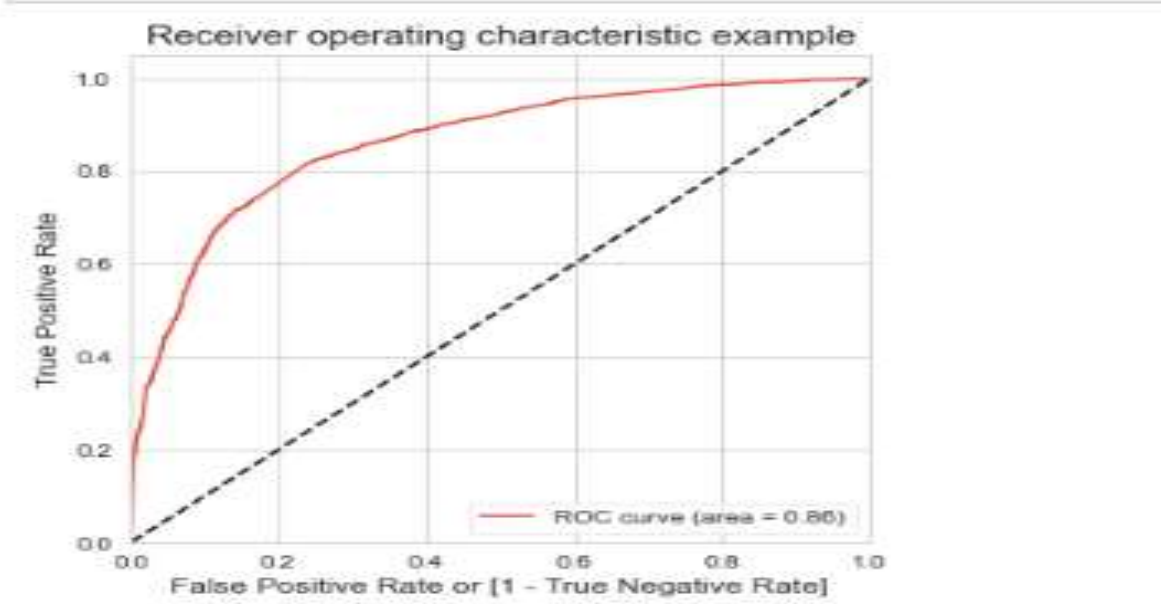
```
Model Accuracy value is             :  78.26 %
Model Sensitivity value is          :  81.67 %
Model Specificity value is          :  76.16 %
Model Precision value is            :  67.86 %
Model Recall value is               :  81.67 %
Model True Positive Rate (TPR)      :  81.67 %
Model False Positive Rate (FPR)     :  23.84 %
Model Poitive Prediction Value is   :  67.86 %
Model Negative Prediction value is  :  87.09 %
```

ANALYSIS

At the cut off of 0.317 below are observations :

- Accuracy : 78.26%
- Sensitivity : 81.67%
- Specificity : 76.16%

ROC CURVE

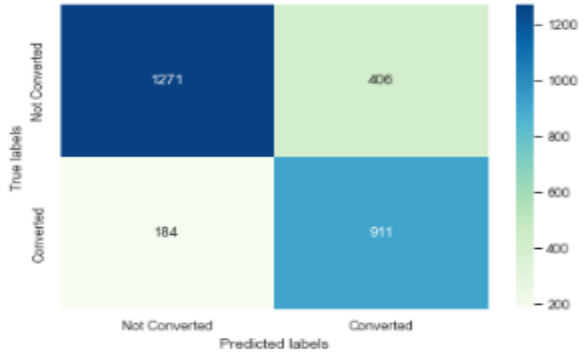


Inference

ROC curve area is 0.86, which indicates that the model is good.

Model Evaluation : Test Dataset

```
In [122]: #Plotting the Confusion Matrix
draw_cm( y_pred_final['Converted_Value'], y_pred_final['final_predicted'], "GnBu")
```




```
In [123]: conf_matrix = confusion_matrix(y_pred_final['Converted_Value'], y_pred_final['final_predicted'])

lg_metrics(conf_matrix)
```

```
Model Accuracy value is              :  78.72 %
Model Sensitivity value is           :  83.2 %
Model Specificity value is           :  75.79 %
Model Precision value is             :  69.17 %
Model Recall value is                :  83.2 %
Model True Positive Rate (TPR)       :  83.2 %
Model False Positive Rate (FPR)      :  24.21 %
Model Poitive Prediction Value is    :  69.17 %
Model Negative Prediction value is   :  87.35 %
```

ANALYSIS

At the cut off of 0.317 below are observations :

- Accuracy : 78.72%
- Sensitivity : 83.2%
- Specificity : 75.79%

RECOMMENDATIONS AND CONCLUSION

- **EDA:**
- SMS messages can have a high impact on lead conversion
- Landing page submissions can help find out more leads
- References for referring a lead can be a good source for higher conversions
- People spending higher than the average time are promising leads, so target them and approaching them can be helpful in conversions.

- **LOGISTIC REGRESSION MODEL:**
- The model shows high sensitivity of 81.67% on the train data set and 83.2% on the test data set
- The model shows high accuracy of 78.26 %
- The ROC curve comes out to be 0.86 which suggests it is a good model
- The model finds correct promising leads which can be converted at a percentage over 80%